Measurement of the Cobb angle on radiographs of patients who have scoliosis. Evaluation of intrinsic error

RT Morrissy, GS Goldsmith, EC Hall, D Kehl and GH Cowie
J Bone Joint Surg Am. 1990;72:320-327.

Measurement of the Cobb Angle on Radiographs of Patients Who Have Scoliosis

Evaluation of Intrinsic Error*

BY RAYMOND T. MORRISSY, M.D.†, GREGORY S. GOLDSMITH, M.D.†, ELMER C. HALL, PH.D.‡, DOUGLAS KEHL, M.D.†, AND G. HENRY COWIE, M.D.†, ATLANTA, GEORGIA

From the Scottish Rite Children's Medical Center, Atlanta

ABSTRACT: **To quantitate the intrinsic error in measurement, fifty anteroposterior radiographs of patients who had scoliosis were each measured on six separate occasions by four orthopaedic surgeons using the Cobb method. For the first two measurements (Set I), each observer selected the end-vertebrae of the curve; for the next two measurements (Set II), the end-vertebrae were pre-selected and constant. The last two measurements (Set III) were obtained in the same manner as Set II, except that each examiner used the same protractor rather than the one that he carried with him. The pooled results of all four observers suggested that the 95 per cent confidence limit for intraobserver variability was 4.9 degrees for Set I, 3.8 degrees for Set II, and 2.8 degrees for Set III. The interobserver variability was 7.2 degrees for Set I and 6.3 degrees for Sets II and III. The mean angles differed significantly between observers, but the difference was smaller when the observers used the same protractor.**

In most scientific endeavors, including medicine, there is a need for accurate measurement. For patients who have scoliosis, the amount of lateral bend is the most important and distinguishing feature of the radiographic examination. The Cobb method[4] is the standard method of quantitating this measurement. Although the Cobb angle is recognized as being a measure of the amount of tilt of the end-vertebrae rather than an objective measure of all aspects of the deformity, it is used to make decisions about the progression of a curve, the need for treatment, and the effectiveness of treatment.

Despite its importance, there is little information about the degree of certainty of changes in the Cobb angle from one radiograph to the next. One study[11] showed the difference between the findings of two examiners to be 3.12 ± 0.48 degrees, and in another[9], the true angle of measurement was estimated to be within ±8.8 degrees, with 95 per cent certainty. Most investigators have considered 5 degrees of change or more between two successive radiographs to be clinically important, even though there is no firm evidence to support the use of this figure[3,8].

In their study of the natural progression of scoliosis, Lonstein and Carlson used a 5-degree difference between the Cobb angles on two successive radiographs as the criterion of progression. Brooks et al. and Rogala et al. also used 5 degrees as the criterion of progression in their epidemiological studies. In recent reports on the evaluation of electrical stimulation in the treatment of scoliosis, 5 degrees was used as indicating progression and thus was the criterion for inclusion in the study[11,12]. In the clinical setting, it is common for practitioners to make recommendations concerning treatment on the basis of an increase in the curve of 5 degrees between two successive radiographs. An increase from 20 to 25 degrees may be a reason to prescribe bracing, and an increase from 40 to 45 degrees may prompt a recommendation for operative treatment.

The purpose of this study was to determine the error, first when the same observer and then when different observers measured the same radiographs. In addition, we attempted to identify the possible sources of error and the relative contribution of each source to the amount of error.

Materials and Methods

Fifty good-quality anteroposterior radiographs of patients who had a thoracic, thoracolumbar, or lumbar scoliosis of between 20 and 40 degrees were selected from the files of the Scottish Rite Medical Center, Atlanta, Georgia. This narrow range of magnitude was selected because when a curve is in this range certain decisions regarding treatment are made and a few degrees of variation in the range are of relatively greater magnitude than in higher ranges of curvature. Two radiographs could not be used because coded numbers on the radiograph apparently had been recorded instead of the measurements of the Cobb angle. In addition, the two measurements were more than seven standard deviations from their respective means and therefore qualified as statistical outliers.

The original radiographs were used to avoid the loss of quality that can result from duplication. Each radiograph was numbered. All radiographs were marked for measure-

* No benefits in any form have been received or will be received from a commercial party related directly or indirectly to the subject of this article. No funds were received in support of this study.

† Scottish Rite Children's Medical Center, 1001 Johnson Ferry Road, N.E., Atlanta, Georgia 30363.

‡ Emory University, 1365 Clifton Road, N.E., Atlanta, Georgia 30322.

TABLE I

SET I: DIFFERENCES BETWEEN ANGLES MEASURED TWICE FOR EACH EXAMINER, WITH THE END-VERTEBRAE SELECTED BY EACH EXAMINER (FREQUENCY AND CUMULATIVE PER CENT DISTRIBUTIONS)

Difference between Angles Measured Twice	Examiner I		Examiner II		Examiner III		Examiner IV		Total	
	No. of Curves	Cumulative Per Cent of Curves	No. of Curves	Cumulative Per Cent of Curves	No. of Curves	Cumulative Per Cent of Curves	No. of Curves	Cumulative Per Cent of Curves	No. of Curves	Cumulative Per Cent of Curves
0 degrees	7	15	9	19	11	23	13	27	40	20.8
1 degree	17	50	13	46	16	56	17	63	63	53.6
2 degrees	11	73	5	56	13	83	15	94	44	76.6
3 degrees	9	92	9	75	2	88	3	100	23	88.5
4 degrees	2	96	5	85	3	94			10	93.8
5 degrees	1	98	2	90	1	96			4	95.8
6 degrees	1	100	2	94	1	98			4	97.9
7 degrees			1	96	1	100			2	99.0
8 degrees			1	98					1	99.5
9 degrees			1	100					1	100.0
Total	48		48		48		48		192	

ment with a soft-lead pencil, and after the measurements had been recorded the markings were erased with trichloracetone.

The radiographs were measured by four examiners who had various levels of experience in managing patients who have scoliosis. The participants included two orthopaedic surgeons who had several years of experience, a fellow in pediatric orthopaedics, and a senior resident who was rotating through the pediatric orthopaedic service. Each examiner measured the radiographs four times, with at least one week between each session. The first set of data (Set I) was derived from forty-eight radiographs that were measured twice by the four examiners. Each examiner recorded his choices of the end-vertebrae and the Cobb angle for each measurement without knowledge of the measurements that he had made one week previously. It was thus possible for the same examiner to select different end-vertebrae on the same radiograph.

The second set of data (Set II) was similarly derived from the same forty-eight radiographs, except that one examiner chose the end-vertebrae that were to be used for all measurements of the Cobb angle. Both Set-I and Set-II measurements were obtained with the instrument that the individual physician carried in his pocket and used daily. Each examiner kept the same instrument for the duration of the study. The third set of data (Set III) was obtained in the same manner as the second set (Set II), except that all examiners used the same protractor. The base-line was inscribed on the protractor, and this base-line, rather than the edge of the protractor, was used for the measurement.

Originally, the study was to include only Set-I data. However, after analyzing the data and sharing the results with the examiners, Set II was designed to assess the observed problems with the Set-I data. The inclusion of Set III was suggested by a reviewer as a means of further assessing the amount of variation in the measuring devices.

Statistical methods: Two complementary approaches to analysis and interpretation were used. The first approach was purely descriptive — that is, it was a simple enumeration of intraobserver differences on successive measurements of the Cobb angle on the forty-eight radiographs. The second approach involved the use of summary statistics from analysis-of-variance calculations to express intraobserver and interobserver variability and to provide 95 per cent prediction limits for the errors in measurement. In addition, as an estimate of the variability in the selection of the vertebrae for the measurements of the Cobb angle, a so-called error index, as defined by Oda et al., was computed for each examiner for the first set of forty-eight duplicate measurements. The number provided a simple means with which to compare the variability of each observer in selecting end-vertebrae for the two readings of the forty-eight radiographs: the larger the number, the more the variability.

Results

Intraobserver Variability

Table I shows the distribution of differences between the pairs of angles that were measured in Set I for each of the four examiners. For example, Examiner I measured seven of the forty-eight curves with perfect agreement (0 degrees of difference between successive measurements) and measured seventeen with 1 degree of difference, and so on. The column for cumulative percentage in Table I indicates that, for Examiner I, 92 per cent of the differences were 3 degrees or less and 98 per cent were 5 degrees or less. For Examiners II, III, and IV, 75, 88, and 100 per cent, respectively, of all pairs differed by 3 degrees or less. Over-all, 88.5 per cent of all pairs differed by 3 degrees or less and 95.8 per cent differed by 5 degrees or less. Conversely, 11.5 per cent of all pairs differed by more than 3 degrees and 4.2 per cent differed by more than 5 degrees. Differences were as large as 9 degrees.

The Set-II data (Table II), for which the selections of the cephalad and caudad vertebrae were fixed, showed some

TABLE II

SET II: DIFFERENCES BETWEEN ANGLES MEASURED TWICE FOR EACH EXAMINER, WITH THE END-VERTEBRAE SELECTED BY ONE EXAMINER (FREQUENCY AND CUMULATIVE PER CENT DISTRIBUTIONS)

Difference between Angles Measured Twice	Examiner I		Examiner II		Examiner III		Examiner IV		Total	
	No. of Curves	Cumulative Per Cent of Curves	No. of Curves	Cumulative Per Cent of Curves	No. of Curves	Cumulative Per Cent of Curves	No. of Curves	Cumulative Per Cent of Curves	No. of Curves	Cumulative Per Cent of Curves
0 degrees	6	17	9	19	13	27	24	50	54	28.1
1 degree	15	48	19	58	21	71	18	88	73	66.1
2 degrees	10	69	15	90	7	85	5	98	37	85.4
3 degrees	10	90	4	98	5	96	1	100	20	95.8
4 degrees	2	94	1	100					3	97.4
5 degrees	2	98			1	99			3	99.0
6 degrees										
7 degrees					1	100			1	99.5
8 degrees									1	100.0
9 degrees	1	100								
Total	48		48		48		48		192	

improvement in intraobserver variability; over-all, 95.8 per cent of all pairs differed by 3 degrees or less and 99 per cent differed by 5 degrees or less. Conversely, 4.2 per cent of all pairs differed by more than 3 degrees and only 1 per cent differed by more than 5 degrees. Again, differences were as large as 9 degrees. Except for Examiner I, intraobserver variability, as expressed by such percentile calculations, improved between Set-I and Set-II measurements.

Set-III data (Table III), for which the cephalad and caudad vertebrae were pre-selected and the same protractor was used, showed that, compared with Set-II data, the agreement between the duplicate measurements for Examiners I and II was improved, but that for Examiners III and IV was not changed much.

Error Index

The error index is a method of expressing the variability in choosing the end-vertebrae on two different occasions (see Appendix). The larger the error index, the more variable was the choice of end-vertebrae between successive measurements of the same radiograph. The error indices that were calculated for the Set-I measurements of Examiners I, II, and III were quite similar. The smaller index for Examiner IV indicated a more consistent choice of end-vertebrae for replicate measurements of the Cobb angle. This apparent consistency for Examiner IV is also reflected in Tables I and II, which show that, among the four examiners, Examiner IV had the least variability in the measurements of the Cobb angle.

TABLE III

SET III: DIFFERENCES BETWEEN ANGLES MEASURED TWICE FOR EACH EXAMINER, USING THE SAME PROTRACTOR, WITH THE END-VERTEBRAE SELECTED BY ONE EXAMINER (FREQUENCY AND CUMULATIVE PER CENT DISTRIBUTIONS)

Difference between Angles Measured Twice	Examiner I		Examiner II		Examiner III		Examiner IV		Total	
	No. of Curves	Cumulative Per Cent of Curves	No. of Curves	Cumulative Per Cent of Curves	No. of Curves	Cumulative Per Cent of Curves	No. of Curves	Cumulative Per Cent of Curves	No. of Curves	Cumulative Per Cent of Curves
0 degrees	9	19	24	50	9	19	19	40	61	32
1 degree	24	69	17	85	23	67	23	88	87	77
2 degrees	7	83	7	100	8	83	6	100	28	92
3 degrees	7	98			6	96			13	98
4 degrees					2	100			2	99
5 degrees	1	100							1	100
Total	48		48		48		48		192	

Mean Angle

The mean angle was the mean of the angles that were measured by one examiner on all of the radiographs for one set of data. The over-all mean angle differed significantly ($p < 0.0001$) among the examiners for Sets I and II, indicating that some examiners consistently tended to measure curves as being larger or smaller than did the other examiners. The only pair of means that was not significantly different (0.05, pairwise Tukey multiple-comparison tests) was that of Examiner II and Examiner III in Set I. The four

TABLE IV

SET I: ERROR INDICES, OVER-ALL MEANS, STANDARD DEVIATIONS, AND NINETY-FIVE PER CENT PREDICTION LIMITS FOR EACH EXAMINER, WITH THE END-VERTEBRAE SELECTED INDEPENDENTLY FOR EACH MEASUREMENT

Characteristics Estimated*	Examiner I	Examiner II	Examiner III	Examiner IV	All†
Error index[1]	0.56	0.63	0.56	0.39	0.54
Over-all mean Cobb angle *(degrees)*	30.7	27.4	27.6	28.9	28.6
S: standard deviation of a measurement of the Cobb angle[2] *(degrees)*	1.6	2.3	1.6	1.0	1.7
S_d: standard deviation of the difference between two successive replicate measurements of the Cobb angle[3] *(degrees)*	2.2	3.3	2.3	1.4	2.4
95% prediction limit for the difference between two successive replicate measurements of the Cobb angle[5] *(degrees)*	4.5	6.7	4.6	3.0	4.9

* See Appendix for footnotes and additional explanation.
† Based on a pooled analysis of variance of the four examiners.

examiners had the same rank order in both sets; Examiner I recorded the largest over-all mean angle, followed by Examiner IV, Examiner III, and Examiner II.

In the Set-III data, for which the same protractor was used, the mean angles were considerably more similar, although they were still significantly different ($p < 0.001$). Examiner I recorded the largest mean angle in all three sets. The other three examiners were in relatively close agreement in all three sets.

Prediction Limit

The intraobserver variability in the measurement of the Cobb angle (Tables I, II, and III) can also be expressed in statistical terms involving three closely related quantities (Tables IV, V, and VI). The standard deviation of a measurement of the Cobb angle and the standard deviation of the difference between two measurements of the Cobb angle are described in the Appendix.

The 95 per cent prediction limit for each examiner indicated the difference between two successive replicate measurements of the Cobb angle that would be exceeded approximately 5 per cent of the time due to an error in measurement. For example, in the Set-I data, for which each examiner selected the end-vertebrae for each measurement of the curve, it could be predicted that the difference between the successive replicate measurements of Examiner II would exceed 6.7 degrees 5 per cent of the time due to an error in measurement. However, in the Set-II data, for which the end-vertebrae were pre-selected, the 95 per cent prediction limit was reduced to 3.4 degrees for Examiner II.

Except for Examiner I, all examiners had some improvement (reduction) in the prediction limit between Set I and Set II. Errors in the measurement of the same radiographs yielded differences in the Cobb angle of as much as 9 degrees. Over-all, approximately 5 per cent of the differences exceeded 4.9 degrees when each examiner selected the end-vertebrae for the measurements of the curve and exceeded 3.8 degrees when the same end-vertebrae were used by all. In Set III, the prediction limit improved for all examiners and, over-all, approximately 5 per cent of the differences exceeded 2.8 degrees.

Intraobserver variability could be alternatively expressed in terms of the percentage of duplicate measurements that differed by 5 degrees or more, on the basis of an assumption that error in the measurements has a normal

TABLE V

SET II: OVER-ALL MEANS, STANDARD DEVIATIONS, AND NINETY-FIVE PER CENT PREDICTION LIMITS FOR EACH EXAMINER, WITH THE END-VERTEBRAE SELECTED BY ONE EXAMINER AND USED FOR ALL MEASUREMENTS

Characteristics Estimated*	Examiner I	Examiner II	Examiner III	Examiner IV	All†
Over-all mean Cobb angle *(degrees)*	30.6	26.8	27.9	29.1	28.6
S: standard deviation of a measurement of the Cobb angle[2] *(degrees)*	1.8	1.2	1.3	0.7	1.3
S_d: standard deviation of the difference between two successive replicate measurements of the Cobb angle[3] *(degrees)*	2.5	1.7	1.9	1.0	1.9
95% prediction limit for the difference between two successive replicate measurements of the Cobb angle[5] *(degrees)*	5.1	3.4	3.9	2.0	3.8

* See Appendix for footnotes and additional explanation.
† Based on a pooled analysis of variance of the four examiners.

TABLE VI

SET III: OVER-ALL MEANS, STANDARD DEVIATIONS, AND PREDICTION LIMITS, WITH USE OF THE SAME PROTRACTOR, WITH THE END-VERTEBRAE SELECTED BY ONE EXAMINER

Characteristics Estimated*	Examiner I	Examiner II	Examiner III	Examiner IV	All†
Over-all mean Cobb angle *(degrees)*	29.8	29.0	29.2	28.5	29.1
S: standard deviation of a measurement of the Cobb angle[2] *(degrees)*	1.2	0.9	1.2	0.7	1.0
S_d: standard deviation of the difference between two successive replicate measurements of the Cobb angle[1] *(degrees)*	1.7	1.0	1.7	1.0	1.4
95% prediction limit for the difference between two successive replicate measurements of the Cobb angle[5] *(degrees)*	3.4	2.0	3.4	2.0	2.8

* See Appendix for footnotes and additional explanation.
† Based on a pooled analysis of variance of the four examiners.

(gaussian) distribution. From Set I, the gaussian estimate was that 4 per cent of the duplicate measurements would differ by 5 degrees or more, due to errors in the measurement. From Set II, the estimate was 1 per cent and from Set III, 0.04 per cent.

Interobserver Variability

Since each examiner measured each radiograph twice in each set, analysis-of-variance calculations could also provide estimates of interobserver variation. It has already been pointed out that the examiners' over-all mean angles differed significantly, implying examiner-to-examiner differences in measuring the sizes of the angles. These differences could be summarized with calculations that were similar to those of Tables IV, V, and VI, but that included examiners as an additional source of variation affecting the 95 per cent prediction limit. The 95 per cent prediction limit, including interobserver variation, was 7.2 degrees for Set I and 6.3 degrees for Sets II and III. Therefore, from the Set-I data, it could be predicted that if two different examiners each measured the angle on the same radiograph and each selected the end-vertebrae, the measurements would differ by more than 7.2 degrees approximately 5 per cent of the time. If, however, the same end-vertebrae were pre-selected for both examiners (Set-II and III data), the measurements of the Cobb angle would differ by more than 6.3 degrees, whether or not the examiners used the same protractor, approximately 5 per cent of the time. The largest difference was 14 degrees; several differences exceeded 10 degrees.

Again, assuming that the errors in measurement were distributed normally, interobserver error could be expressed in terms of the percentage of the angles, each measured by two examiners, that would differ by 5 degrees or more. From the Set-I data, the gaussian estimate was that 16 per cent of such measurements would differ by 5 degrees or more and from Set-II and III data, that 11 per cent would.

Identifiable Causes of Error

In each instance in which the measurements of two examiners differed by more than 6 degrees, the radiographs were evaluated qualitatively. All of the curves were either thoracolumbar or lumbar. Also, the end-plate of one or both of the end-vertebrae appeared indistinct, cup-shaped, or biconcave, due to the tilt of a vertebral body that was in lordosis or kyphosis.

In an effort to determine why the measurements of the mean angles differed among the examiners, the protractors were studied after Sets I and II were completed. Except for Examiner IV, who used the protractor on a Pedroli ruler (a ruler on which the base-line is inscribed), the examiners

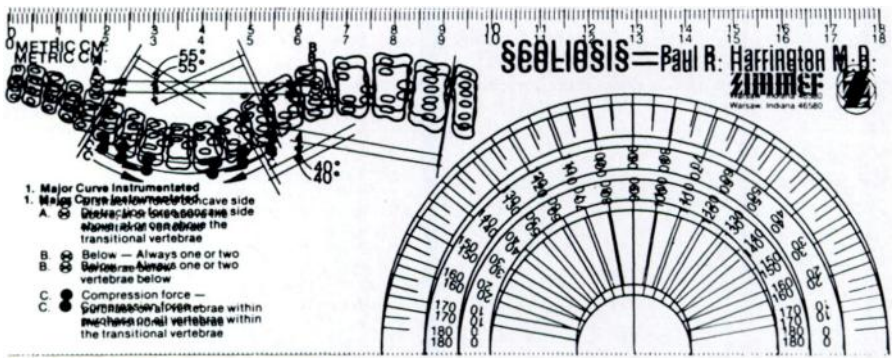

FIG. 1

The protractors of Examiners I and II, superimposed. The bottom edge of each ruler differs in relation to the angles. The smaller the angle, the greater the difference.

used standard protractors, distributed by various companies. The instruments varied widely. Figure 1 shows the protractors that were used by Examiners I and II, whose overall measurements of the mean angles differed the most. The smaller the angle, the larger the error, while at 90 degrees the lines of the two protractors are superimposed. The reason for this appears to be that, in some protractors, the baseline is the bottom of the ruler rather than an absolute line that is related to the angles. Therefore, if more than the correct amount of plastic is removed from the bottom of the ruler when the ruler is stamped, it will measure a larger angle.

Discussion

Several authors have attempted to determine the accuracy of the Cobb measurement. Understanding and interpreting the data in these reports is difficult, since it is often not clear how the numbers were derived[9], what the numbers refer to (standard error of the mean or standard deviation)[1], or how the authors' interpretation relates to intraobserver or interobserver differences[6].

Sevastikoglou and Bergquist found the interobserver error for the Cobb method to be 3.12 ± 0.48 degrees between two examiners. They did not quantitate the intraobserver error. Beekman and Hall studied a series of single measurements of radiographs of curves that were between 5 and 25 degrees. Two physicians each selected the end-vertebrae and measured the radiographs once. The mean difference between the measurements of the two examiners was 4.2 ± 0.95 degrees, and the range was 1 to 10 degrees. Beekman and Hall also did not study the intraobserver error. Gross et al. studied the errors in measurement of three examiners who measured each radiograph five times. The magnitude of the curves ranged from 4.5 to 106.5 degrees. The 95 per cent confidence limit for the intraobserver error of the three examiners was reported to be between ±2 and ±4 degrees. The interobserver error when each examiner used five measurements was 2.3 ± 1.0 degrees.

Oda et al. studied the intraobserver and interobserver error among five orthopaedic surgeons who measured fifty radiographs of scoliotic curves twice, selecting the end-vertebrae in the manner that was used in this study to obtain the Set-I data. The intraobserver error for duplicate measurements of the same radiograph averaged 12.61 degrees. The 95 per cent confidence interval for the five orthopaedists ranged from 10.0 to 30.2 degrees. The variability in the selection of the end-vertebrae was the largest factor in the error in repeated measurements of the same radiograph. The interobserver error among the five orthopaedists averaged 20.18 degrees. The authors calculated that there was a 95 per cent certainty that a single reading by any orthopaedist would be within 8.8 degrees of the true angle.

Our study was designed to determine the intraobserver and interobserver error under circumstances that mirrored reality. A physician who sees a patient who has scoliosis should know the precision and reproducibility of the Cobb angle that he or she measures. Similarly, if different physicians (for example, residents) will be measuring successive radiographs of a patient, this interobserver error should be known. In research studies, it is important to know how much error is introduced when different examiners measure radiographs of different patients.

In the current study, the improvement that was obtained when the end-vertebrae were constant can be seen by comparing Tables I and II. When selection of different vertebrae was permitted, 4.2 per cent of the measurements of all Cobb angles by the four examiners varied by more than 5 degrees. When the end-vertebrae were constant, only 1 per cent of the measurements of all Cobb angles varied by more than 5 degrees. The importance of selecting the same end-vertebrae can be seen by comparing the results of Examiners II and IV. When each examiner selected the end-vertebrae, only 75 per cent of the duplicate measurements of Examiner II were within 3 degrees of each other, while for Examiner IV the value was 100 per cent. When the end-vertebrae were pre-selected and constant, 98 per cent of the measurements of Examiner II were within 3 degrees of each other, illustrating that selection of different end-vertebrae was the largest source of error.

The examiners in our study were much better at selecting the same end-vertebrae than were the examiners in the study by Oda et al. This is reflected by the error index, which ranged between 0.39 and 0.63 for the four examiners in our study and between 0.8 and 2.4 in the study by Oda et al. The reason for this large discrepancy is unclear. All but one of the examiners in the study of Oda et al. used a grease marker, and the one examiner who used a pencil had the smallest standard deviation. All of the measurements in our study were done with a soft-lead pencil designed for marking radiographs.

When the variability of selection of the end-vertebrae was eliminated, the amount of actual error in the measurements among the examiners was relatively small. Although Examiner I appeared to be the least precise with the protractor, only five of forty-eight replicate measurements varied by more than 3 degrees and only one, by more than 5 degrees. When the end-vertebrae were constant for all examiners, the error was 6.3 degrees, but when each examiner was permitted to select the end-vertebrae, the error was 7.2 degrees. Both of these values are larger than the largest intraobserver error. This confirms that the selection of end-vertebrae is the single largest source of intraobserver error, and that multiple examiners introduce another source of error in addition to the intraobserver error and the selection of end-vertebrae.

In actual clinical practice, curves may change, and additional vertebrae may be added to the curve. However, this study of intraobserver error was not aimed at establishing the true severity of the curve, but only the reproducibility of the examiners' measurements. It is of interest that the selection of end-vertebrae, something that might be assumed to need clinical judgment, was not correlated with the level of experience of the examiner. This was true also for the intraobserver error.

All of these findings emphasize the need to note the end-vertebrae of the curve accurately, to measure previous radiographs again if the end-vertebrae appear to have changed, and to have one person measure all of a patient's radiographs whenever possible.

One additional factor must be considered when evaluating each set of data and the improvement over the three sets of data. All examiners were aware of the purpose of their measurements and of the results of their performance after each set and before performing the measurements in the next set. The conversation among the examiners after each set was analyzed, and there is no doubt that some were sensitized to their performance and that they competed among themselves to improve their precision. In this sense, the study does not simulate a real-life situation; it is likely that the results in clinical practice are less precise. Our study also shows that precision can be improved through awareness and, perhaps, practice.

The variation in the mean angles was at first thought to be due to consistent idiosyncrasies in the techniques for measurement. However, it is likely that it was due to the variations among the protractors. The error that was introduced by the protractor, however, was not sufficient to change the interobserver error when the second decimal was rounded off. Nevertheless, care should be taken to select an accurate instrument for measurement. A protractor that has a clearly inscribed base-line is essential. Protractors in which the cut edge is used as the base-line may be inaccurate, vary from one to another, and result in falsely high measurements, as was the case for Examiner I.

It is important to emphasize that this study quantitated only intrinsic error. In the clinical setting, two radiographs made at different times are measured. This allows for the introduction of extrinsic error, for which the three main sources are the position of the patient, the position of the radiographic tube, and the time of day when the radiograph was made[5,11,12].

Our data indicate that, under the best of circumstances, when a careful examiner uses a lead pencil and measures the same end-vertebrae, with no extrinsic error, there is a 95 per cent chance that the error in measurement will be less than 3 degrees. Some examiners will be better and some will be worse, and each may determine his or her own precision. In critical situations, the average of two or more measurements on each radiograph is a method of reducing error. In patients who have scoliosis, decisions for treatment — especially those involving an operation — are usually not urgent. This provides the opportunity to determine the change in the Cobb angle among three radiographs that have been made three to four months apart, instead of between only two radiographs. Three successive radiographs on which the Cobb angle is 23, 26, and 29 degrees have a far different meaning than three in which it is 23, 28, and 25 degrees.

If more than one examiner make the measurements in a clinical study, the interobserver error should be determined, because even under the most ideal circumstances the difference in 5 per cent of the measured angles will exceed 6.3 degrees. The usual criterion of 5 degrees of change between two successive radiographs may not be adequate to determine progression. Finally, the instruments that are used for drawing on the radiographs and for measurement should be accurate and, in multicenter studies, standardized.

Appendix

1. The error index equals

$$\frac{\sum[(U_1 - U_2)^2 + (L_1 - L_2)^2]^{1/2}}{48}$$

where U_1,U_2,L_1, and L_2 are the first and second choices of the cephalad and caudad end-vertebrae for successive measurements of the Cobb angle. Note that an examiner who always selected the same cephalad and caudad vertebrae for successive measurements would have an error index of zero. For any given pair of successive measurements for which the end-vertebrae are independently selected, the quantity

$$[(U - U_2)^2 + (L_1L_2)^2]^{1/2}$$

is the euclidian distance between the two points (U_1,L_1) and (U_2,L_2) plotted on a plane with "upper" (U) and "lower" (L) axes.

2. S is the square root of the within-pairs mean square from a one-way analysis of variance of each examiner's forty-eight pairs of measurements. It is an estimate of the standard deviation of replicate measurements of the Cobb angle, which expresses the intrinsic error in the measurement for each examiner. For example, in Set I, multiple measurements of a given angle by Examiner II had an estimated standard deviation of 2.3 degrees. This was the largest standard deviation among the four examiners, and it was associated with the largest error index. Examiner IV had the smallest standard deviation and the smallest error index. Except for Examiner I, the standard deviations were smaller in Set-II measurements than in Set-I measurements. Set-III standard deviations were the smallest for all examiners.

3. For two independent successive measurements of the Cobb angle of the same radiograph, the variance of the difference between the two measurements is twice the variance of a single measurement — that is, if S^2 is the estimated variance of a single measurement, $2S^2$ is the estimated variance of the difference between two measurements. Hence, the standard deviation of the difference between two successive measurements is $(2S^2)^{1/2}$ or S_d, which expresses the variability of differences between two successive measurements of the Cobb angle on the same radiograph. In other words, if the directional differences (first measurement minus second measurement) are recorded for all pairs of measurements of the Cobb angle, these quantities provide an estimate of the standard deviations of such differences.

4. Ninety-five per cent prediction limits are obtained from the formula

$$\mu_d \pm t_{1-\alpha/2,n-1}s_d \left[1 + \frac{1}{n}\right]^{1/2}$$

where μ_d is known to be zero (that is, the average difference between all possible pairs of successive replicate measurements of the Cobb angle, first measurement minus second measurement, is zero), n is the number of pairs of measurements (forty-eight in this study), and $t_{1-\alpha/2,n-1}$ is the $1-\alpha/2$ fractile of a t distribution with $n-1$ degrees of freedom (for $\alpha = 0.05$ and $n = 48$, $t = 2.01$). For this particular example, the expression after the $\pm$ symbol in the formula reduces to 2.03 S_d. Hence, differences between successive replicate measurements will exceed 2.03 S_d degrees approximately 5 per cent of the time.

References

1. BEEKMAN, C. E., and HALL, VIVIAN: Variability of Scoliosis Measurement from Spinal Roentgenograms. Phys. Ther., **59:** 764-765, 1979.
2. BROOKS, H. L.; AZEN, S. P.; GERBERG, E.; BROOKS, R.; and CHAN, L.: Scoliosis: A Prospective Epidemiological Study. J. Bone and Joint Surg., **57-A:** 968-972, Oct. 1975.
3. BROWN, J. C.; AXELGAARD, JENS; and HOWSON, D. C.: Multicenter Trial of a Noninvasive Stimulation Method for Idiopathic Scoliosis: A Summary of Early Treatment Results. Spine, **9:** 382-387, 1984.
4. COBB, J. R.: Outline for the Study of Scoliosis. *In* Instructional Course Lectures, The American Academy of Orthopaedic Surgeons. Vol. 5, pp. 261-275. Ann Arbor, J. W. Edwards, 1948.
5. DAWSON, E. G.; SMITH, R. K.; and MCNIECE, G. M.: Radiographic Evaluation of Scoliosis. A Reassessment and Introduction of the Scoliosis Chariot. Clin. Orthop., **131:** 151-155, 1978.
6. GROSS, CLIFFORD; GROSS, MICHAEL; and KUSCHNER, STUART: Error Analysis of Scoliosis Curvature Measurement. Bull. Hosp. Joint Dis. Orthop. Inst., **43:** 171-177, 1983.
7. LONSTEIN, J. E., and CARLSON, J. M.: The Prediction of Curve Progression in Untreated Idiopathic Scoliosis during Growth. J. Bone and Joint Surg., **66-A:** 1061-1071, Sept. 1984.
8. MCCOLLOUGH, N. C., III: Nonoperative Treatment of Idiopathic Scoliosis Using Surface Electrical Stimulation. Spine, **11:** 802-804, 1986.
9. ODA, MARJORIE; RAUH, STEPHEN; GREGORY, P. B.; SILVERMAN, F. N.; and BLECK, E. E.: The Significance of Roentgenographic Measurement in Scoliosis. J. Pediat. Orthop., **2:** 378-382, 1982.
10. ROGALA, E. J.; DRUMMOND, D. S.; and GURR, JEAN: Scoliosis: Incidence and Natural History. A Prospective Epidemiological Study. J. Bone and Joint Surg., **60-A:** 173-176, March 1978.
11. SEVASTIKOGLOU, J. A., and BERGQUIST, E.: Evaluation of the Reliability of Radiological Methods for Registration of Scoliosis. Acta Orthop. Scandinavica, **40:** 608-613, 1969.
12. ZETTERBERG, C.; HANSSON, T.; LIDSTRÖM, J.; IRSTRAM, L.; and ANDERSSON, G.: Daytime Postural Changes of the Scoliotic Spine. Orthop. Trans., **7:** 7-8, 1983.