International Scholarly Research Network
ISRN Rheumatology
Volume 2011, Article ID 571698, 8 pages
doi:10.5402/2011/571698

Research Article

Optimizing Reliability of Digital Inclinometer and Flexicurve Ruler Measures of Spine Curvatures in Postmenopausal Women with Osteoporosis of the Spine: An Illustration of the Use of Generalizability Theory

Norma J. MacIntyre, Lisa Bennett, Alison M. Bonnyman, and Paul W. Stratford

School of Rehabilitation Science, McMaster University, IAHS, Room 403, 1400 Main Street West, Hamilton, ON, Canada L8S 1C7

Correspondence should be addressed to Norma J. MacIntyre, macint@mcmaster.ca

Received 1 December 2010; Accepted 2 January 2011

Academic Editors: A. Adebajo and K. Uusi-Rasi

The study illustrates the application of generalizability theory (G-theory) to identify measurement protocols that optimize reliability of two clinical methods for assessing spine curvatures in women with osteoporosis. Triplicate measures of spine curvatures were acquired for 9 postmenopausal women with spine osteoporosis by two raters during a single visit using a digital inclinometer and a flexicurve ruler. G-coefficients were estimated using a G-study, and a measurement protocol that optimized inter-rater and inter-trial reliability was identified using follow-up decision studies. The G-theory provides reliability estimates for measurement devices which can be generalized to different clinical contexts and/or measurement designs.

1. Introduction

Measuring devices are used routinely in rheumatology clinical examinations and research. Reliability analysis quantifies the consistency of examinee performance [1, 2]. When differences arise among repeated measurements performed on a truly stable examinee, it is attributed to measurement error. Not surprisingly, the clinical literature has devoted considerable attention to reliability studies to ensure that the measurements obtained are reliable [3–5]. It is important to know the magnitude of the error variance for a given measurement in order to determine the confidence in a measured value and to assess change in the examinee over time. Applied in the context of reliability analysis, measurement error is an all-encompassing term that includes inherent variation in the examinee, inconsistencies within and between raters, and many other sources of potential variation excluding true differences among examinees under investigation. Typically, two coefficients—the reliability coefficient and the standard error of measurement (SEM)—are used to characterize the reliability of a measure [1]. The reliability coefficient is a unitless quantity. As such, it comments on the relative, reliability of a measure. The SEM is a measure of absolute reliability in that it expresses measurement error in the same units as the original measurement. For a measure to be clinically useful it must possess a sufficiently high reliability coefficient and a sufficiently low SEM. Despite the availability of reliability studies, it is challenging to select information applicable to a particular clinical context.

It is recommended that each clinical setting establish reliability for measurements obtained by their specific assessors (raters) on their particular patient population. This position is held, in part, because studies of measurement error in the clinical arena have predominantly adopted a classical test theory (CTT) framework [1, 2]. CTT states that the observed score (i.e., the measured value) is equal to the sum of the true score and measurement error. The true score is conceptualized as the average score that would be obtained from an infinite number of measurements performed on a truly stable examinee. Consistent with this conceptualization is that the distribution of measurement error values would represent a normal distribution with a

mean of zero. CTT also dictates that true scores and error scores are independent. CTT defines the reliability coefficient (R) as the ratio of true score variance to observed score variance (i.e., sum of true and error variances)

$$\mathrm{R} = \frac{\text{true score variance } (\sigma_t^2)}{\text{observed score variance } (\sigma_x^2)},$$
$$\mathrm{R} = \frac{\text{true score variance } (\sigma_t^2)}{\text{true score variance } (\sigma_x^2) + \text{error variance } (\sigma_e^2)}. \quad (1)$$

The SEM is equal to the square-root of the error variance. The variance terms are obtained from a one-way analysis of variance (ANOVA). Finally, because measurements take place in context, measurement properties comment on the inextricable link among measure, examinees, and measurement process: tests and measures do not have reliabilities, while the measures' scores do [2].

Despite the common use of CTT for characterizing reliability, there are several limitations. First, the term "true" score can be confusing on several counts. When applied in a reliability context, the true score does not comment on the extent to which a measure assesses what it is intended to measure (i.e., its meaning when applied in a validity context). Also, an examinee may have different true scores depending on the study design. For example, the apparent true score for an examinee may be different for an inter-rater study design compared to a inter-trial study design. A second limitation concerns the interpretation of the error term. Although in theory it represents random measurement error, there is no way of distinguishing whether this assumption is true. Furthermore, like the true scores, it is likely that the magnitude of measurement error will be different for different study designs. Finally, CTT does not provide a coherent method for optimizing a measurement process. For example, an investigator might be interested in determining whether a greater gain in reliability could be achieved by increasing the number of raters or by increasing the number of assessments by a single rater. Applying CTT, the investigator would conduct two studies. For the results of each study, the investigator could apply the Spearman-Brown prophecy formula to estimate the impact of altering the number of raters or the number of trials. However, there is no elegant method for combining the results from these studies to determine whether it is better to increase the number of raters or to increase the number of trials. Collectively, these shortcomings led to the development of generalizability theory (G-theory) [6].

G-theory differs from CTT as summarized in Table 1 and builds on it in the following ways [7]. Rather than focusing on a "true" score, the G-theory comments on a "universe" score. The universe score represents the mean score for an examinee over all conditions of interest to the clinician/investigator. These conditions define the universe of admissible observations. The term "facet" is used to describe the conditions of measurement. Thus, in the previous example, the universe of admissible observations includes raters and trials. The term "population" is used to describe the objects of measurement. Having identified the population and facets of interest, the next step is to conduct a study to estimate the variance components. Within the G-theory lexicon, this is referred to as a generalizability study (G-study). Numerous G-study designs exist [8] and it is beyond the scope of this monograph to provide a review of each. Accordingly, for illustrative purpose we will restrict our commentary to a fully crossed design that is frequently reported and of interest to clinicians and investigators. For a fully crossed design, all objects of measurement are assessed by all levels of all facets. Once again, suppose the universe of admissible observations consisted of raters and trials. An investigator conducted a study where two raters each performed three trials on all of the objects of measurement (patients). This fully crossed design is represented as "patients X raters X trials". Seven sources of variance can be identified from this study design: patients (σ_p^2), raters (σ_r^2), trials (σ_t^2); the two-way interaction of patients and raters (σ_{pr}^2), patients and trials (σ_{pt}^2), raters and trials (σ_{rt}^2); the three-way interaction of patients and raters and trials (error, σ_{prt}^2). These variance components can be used to calculate generalizability coefficients (G-coefficients) that are roughly equivalent to R. The equivalent G-coefficient for an inter-rater reliability is

$$\mathrm{G}_{\text{inter-rater}} = \frac{\left(\sigma_p^2 + \sigma_t^2 + \sigma_{pt}^2\right)}{\left(\sigma_p^2 + \sigma_r^2 + \sigma_t^2 + \sigma_{pr}^2 + \sigma_{pt}^2 + \sigma_{rt}^2 + \sigma_{prt}^2\right)}, \quad (2)$$

and the equivalent G-coefficient for inter-trial reliability is

$$\mathrm{G}_{\text{inter-trial}} = \frac{\left(\sigma_p^2 + \sigma_r^2 + \sigma_{pr}^2\right)}{\left(\sigma_p^2 + \sigma_r^2 + \sigma_t^2 + \sigma_{pr}^2 + \sigma_{pt}^2 + \sigma_{rt}^2 + \sigma_{prt}^2\right)}. \quad (3)$$

Having identified the variance components from a single G-study, the investigator would then apply these results to guide decisionmaking concerning the optimal measurement strategy. This type of study is referred to as a Decision study (D-study). A D-study is similar to applying the Spearman-Brown prophecy formula; however, with a D-study it is possible to examine the impact of varying the number of raters and number of trials simultaneously.

Table 1: Comparison of differences between classical test theory and generalizability theory.

Classical test theory	Generalizability theory
True score	Universe of admissible observations' score
One identifiable source of "error" variance	Multiple sources of identifiable "error" variances
One-way ANOVA	Factorial ANOVA
"What if" optimizing assessment method: Spearman Brown	"What if" optimizing assessment method: design study

2. Examplar Application of G-Theory

In our clinical research setting, we were interested in designing a study involving measurement of spine curvatures

in postmenopausal women with osteoporosis. Women with osteoporosis are susceptible to deformities in the axial skeleton including hyperkyphosis and flattened or accentuated lumbar lordosis [9]. Clinical practice guidelines for rehabilitation of women with spine osteoporosis include postural assessment and correction of abnormal spinal curvatures [10]. The American Physical Therapy Association Section on Geriatrics recommends measuring kypholordosis using a surveyor's flexicurve ruler [11]. Measuring change in kyphosis is important since hyperkyphosis is associated with increased spinal loads which increase the risk for subsequent fracture [12], and women with a kyphotic index ≥ 13 have reduced cardiovascular fitness, muscle strength, and physical function [13, 14]. Although less studied, assessment of lumbar lordosis is also important in this patient group given that prescription of certain orthoses (e.g., the PTS brace) is contraindicated in those with flattened lordotic curvatures due to the loads imparted to this region of the axial skeleton. Thus, reliable measurement of spine curvatures aids in the classification of women with postmenopausal osteoporosis at increased risk for fracture, prescription of appropriate bracing, and ongoing monitoring of progression and response to therapeutic interventions aimed to improve abnormal postures. To plan our future study, a pilot study was needed to evaluate and optimize the reliability of values obtained using two common clinical methods for assessing spine curvatures.

Therefore, our purpose was to illustrate the application of the tools of the G-theory to investigate the inter-trial and inter-rater reliability of spine curvature measures in postmenopausal women with osteoporosis of the spine using two common methods—the digital inclinometer and the flexicurve ruler, in order to establish an optimal measurement protocol. For comparison, the inter-trial and inter-rater reliability of these measures were also determined using CTT.

3. Methods

3.1. Participants. Nine women were recruited through a local osteoporosis clinic. Women were eligible for inclusion in the study if they were 60 years of age or older, were postmenopausal (self-reported absence of menses for more than 1 year), were clinically diagnosed with osteoporosis by a physician, and had a history of one or more vertebral fracture. Participants were excluded from the study if they were not community ambulators, had cognitive difficulties, were unable to understand written or spoken English, or had a vertebral fracture within three months prior to commencement of the study. The study protocol was approved by our institutional Research Ethics Review Board, and all participants provided written informed consent prior to the start of the study.

3.2. Spine Curvature Measurements. During a single visit, spine curvatures were measured by two raters using two different measurement devices. Clothing covering the back and footwear were removed to ensure accurate identification of bony landmarks and consistent standing posture. Participants were instructed to stand erect and maintain their best posture throughout the procedure. Each rater followed a standardized protocol to acquire triplicate measurements using the digital inclinometer and the flexicurve ruler.

3.2.1. Digital Inclinometer. A digital inclinometer (Saunder's digital inclinometer, Empi Therapy Solutions) was used according to the manufacturer's recommended procedure [15] to measure joint angle at the cervicothoracic, thoracolumbar, and lumbosacral junctions as described here in brief. The arch attachment was fixed to the inclinometer, and the rater held this portion of the inclinometer when zeroing the instrument and taking all measurements. The following three landmarks were palpated and marked with small, circular stickers: the C7-T1 interspace (CT), the T12-L1 interspace (TL), and the sacral midpoint from which the lumbosacral interspace (LS) was identified approximately 3.0 cm superiorly. After landmarking, the inclinometer was placed on a flat vertical surface and the digital reading was set to zero degrees. The inclinometer was initially placed at CT, the angle was then read and recorded by a third person, and the inclinometer was zeroed; the inclinometer was placed at TL, the angle was read and recorded by a third person, the inclinometer was zeroed, and the inclinometer was placed at LS, the angle was read and recorded by a third person. The entire measurement procedure was repeated three times in a row by each of the two raters who were blinded to the results.

3.2.2. Flexicurve Ruler. A 61-cm long flexicurve ruler (Arts Supply Store, Hamilton, ON) was used according to the instructional CD distributed by the American Physical Therapy Association Geriatrics Division [11]. The spinous process of the seventh cervical vertebra (C7) and the LS interspace were palpated and marked with small, circular stickers. The flexicurve ruler was molded along the participant's spine, making sure the shape of the thoracic and lumbar curves was retained and that there were no spaces between the participant's skin and flexicurve ruler. Marks were placed on the flexicurve ruler to correspond with the C7 mark superiorly and the LS interspace mark inferiorly. The flexicurve ruler was carefully removed from the participant's spine and placed onto plain white graph paper. The participant's study identification number, date, and measurement number were recorded at the top of the graph paper. The C7 spinous process and LS interspace marks on the ruler were placed along the same vertical line. The side of the flexicurve ruler that was contacting the participant's skin was traced onto the paper. After tracing the spine curvature on the graph paper, the flexicurve ruler was straightened and the flexicurve ruler procedure was repeated three times in a row by each rater.

The traced curves were landmarked such that a vertical line was drawn to connect the C7 mark (most superior point), and the LS interspace mark (most inferior point) and a perpendicular line was drawn at the TL level. For each trial, KI was calculated according to the following formula:

$$\text{KI} = \frac{(\text{thoracic width} \times 100)}{(\text{thoracic length})}, \tag{4}$$

where thoracic width is the greatest width from the thoracic curve to the vertical line and thoracic length is the distance from the C7 mark to the junction of the thoracic and lumbar curves.

For each trial, LI was calculated according to the following formula:

$$\text{LI} = \frac{(\text{lumbar width} \times 100)}{(\text{lumbar length})}, \quad (5)$$

where lumbar width is the greatest width from the lumbar curve to the vertical line joining C7 and the LS interspace, and lumbar length is the distance from the junction of the thoracic and lumbar curves to the LS interspace.

3.3. Raters. The raters, an undergraduate student with no prior experience using either method of measurement and a physiotherapist with minimal prior experience using a digital inclinometer and no prior experience with the flexicurve ruler, received brief training. The user's manual for the digital inclinometer [15] was studied, and an instructional CD on how to use a flexicurve ruler to measure spine curvatures [11] was viewed by each tester. Practical experience was gained by completing the measurement protocols during two mock trials prior to the start of the pilot study.

3.4. Statistical Analyses. Descriptive statistics were calculated using SPSS v18 (www.spss.com). G-theory was applied using G_String_III version 5.4.2 for Windows [16]. First, a G-study was completed to estimate G-coefficients for the overall variation that can be attributed to the sources of variation (called facets which in this case are the patients, the trials and the raters) and their interactions and the proportions of variation attributed to trials and raters. Follow-up D-studies were performed to identify the optimal measurement protocol for obtaining reliable measures of spine curvatures by varying the number of raters and the number of trials. G-coefficients ≥ 0.80 were considered desirable. For comparison, CTT was also applied. Inter-trial reliability was determined for each rater based on variance components for between- and within-subject factors, and the average of the two values is reported. Inter-rater reliability was determined for each trial based on variance components for between- and within-subject factors and the average of the three values is reported. Absolute reliability of each spine curvature measure was also determined as the standard error of the measurement (SEM) calculated as the square-root of the mean square estimate for the error term determined using G-theory and CTT.

4. Results

The characteristics of the patients are summarized in Table 2. The average spine curvature measures acquired by each of the raters are shown in Table 3. Six of the nine women in our convenience sample exceeded the clinical cutpoint for hyperkyphosis (KI $\geq$ 13) according to measures acquired by at least one of the raters. All women were living independently in the community.

Table 2: Characteristics of 9 postmenopausal women with osteoporosis of the spine.

Variable	Mean (SD)	Minimum, maximum
Age (years)	71.6 (8.9)	63, 76
Height (cm)	156.1 (8.7)	147.2, 162
Weight (kg)	71.2 (24.2)	59.4, 94
Cervicothoracic angle (degrees)[a]	36.1 (9.99)	17.5, 49.2
Thoracolumbar angle (degrees)[a]	51.4 (13.72)	27.2, 72.0
Lumbosacral angle (degrees)[a]	31.9 (9.17)	15.0, 50.2
Kyphotic Index[b]	13.2 (5.07)	5.8, 19.5
Lordotic Index[b]	13.9 (3.22)	9.0, 18.2

[a] calculated as mean of the average values acquired by each of the two raters for each subject using the digital inclinometer.
[b] segment width $\times$ 100/segment length; calculated as mean of the average values acquired by each of the two raters for each subject using the flexicurve ruler.

Table 4 compares and contrasts the estimates of variance components that are determined for measures of KI using G-theory and CTT. Both methods partition variance due to patients, however, the error variance in CTT includes other sources of variance depending upon the measurement design. When assessing inter-rater reliability, the error variance component includes variance due to trial. When assessing inter-trial reliability, the error variance component includes variance due to rater.

Table 5 shows that the estimates of inter-trial and inter-rater reliability of the spine curve measures are comparable whether using G-theory or the CTT. The inter-trial reliability was high for all measures and inter-rater reliability was greatest for KI.

Data from the G-study were used to establish a reliable measurement protocol through D-studies. Figures 1(a) and 1(b) illustrate how the inter-trial reliability changes with increasing numbers of trials when varying numbers of raters perform the measures. For a given rater, all measures are reliable. Minimal gains in reliability are achieved when performing more than 1 trial using the digital inclinometer (Figure 1(a)) and when performing more than 3 trials using the flexicurve ruler (Figure 1(b)). Figures 1(c) and 1(d) illustrate how the inter-rater reliability changes with increasing numbers of raters when different numbers of trials are performed. Measures of CT, TL, and LS angle have acceptable reliability when measured by 5, 2, and 3 raters, respectively, and there is minimal improvement in reliability when more than 1 trial is completed by each rater (Figure 1(c)). By comparison, measures of KI and LI acquired in duplicate by two raters have acceptable reliability (Figure 1(d)).

Table 3: Mean (SD) spine curvature values over 3 trials acquired by 2 raters in 9 women with spine osteoporosis.

Patient	Cervicothoracic angle[a]		Thoracolumbar angle[a]		Lumbosacral angle[a]		Kyphotic index[b]		Lordotic index[b]	
	Rater 1	Rater 2	Rater 1	Rater 2	Rater 1	Rater 2	Rater 1	Rater 2	Rater 1	Rater 2
1	51.0 (2.6)	41.7 (0.6)	77.3 (1.5)	66.7 (4.0)	34.0 (2.6)	28.3 (4.9)	16.5 (0.5)	16.8 (1.6)	13.9 (0.2)	11.4 (3.3)
2	48.7 (12.7)	16.8 (1.0)	36.0 (11.3)	27.0 (2.0)	44.3 (4.0)	39.7 (2.1)	6.6 (0.6)	47.0 (1.7)	19.6 (1.0)	7.4 (0.7)
3	41.0 (0.0)	22.7 (0.6)	47.7 (1.2)	47.3 (0.6)	20.7 (1.5)	38.7 (1.2)	12.3 (1.1)	12.9 (0.4)	9.4 (1.2)	10.2 (0.5)
4	18.7 (2.1)	16.3 (1.2)	28.7 (3.2)	25.7 (2.1)	30.0 (0.0)	31.7 (1.2)	5.7 (0.7)	5.9 (1.1)	11.7 (0.4)	12.3 (1.2)
5	42.0 (3.5)	42.3 (1.5)	51.0 (5.0)	65.7 (1.5)	22.0 (3.6)	36.0 (2.0)	15.6 (1.7)	18.4 (1.6)	17.1 (2.3)	17.0 (1.4)
6	42.7 (3.8)	55.7 (1.5)	55.7 (3.5)	77.0 (2.6)	31.0 (2.0)	33.7 (1.5)	18.6 (0.5)	20.4 (1.1)	16.3 (0.7)	14.4 (0.7)
7	28.7 (1.5)	34.7 (2.1)	39.0 (1.0)	44.7 (2.1)	16.0 (1.7)	14.0 (1.0)	8.8 (1.4)	7.6 (0.8)	8.0 (2.3)	9.9 (1.4)
8	28.3 (0.6)	40.0 (1.0)	45.3 (0.6)	57.0 (1.7)	33.0 (1.0)	29.0 (2.0)	13.4 (0.8)	15.0 (0.6)	13.5 (1.0)	16.3 (0.7)
9	39.0 (3.6)	47.0 (2.0)	54.0 (3.6)	59.3 (3.1)	38.0 (2.6)	38.0 (1.0)	16.2 (0.6)	19.3 (1.6)	15.8 (0.9)	16.5 (1.2)

[a] measured using digital inclinometer, degrees
[b] measured using flexicurve ruler.

Table 4: Estimates of variance components[a] for Kyphotic index using G-theory and classical test theory.

Variance component	G-theory σ^2	Classical Test Theory σ^2				
		Rater 1	Rater 2	Trial 1	Trial 2	Trial 3
Patient	25.263	21.227	30.303	23.593	25.733	25.233
Rater	0.488	—	—	—	—	—
Trial	0.083	—	—	—	—	—
Patient ∗ rater	0.563	—	—	—	—	—
Patient ∗ trial	0	—	—	—	—	—
Rater ∗ trial	0.098	—	—	—	—	—
Error	1.023	0.919	1.256	1.901	2.974	1.641

[a] estimates having negative values are set to zero.

Table 5: Reliability of spine curvature measures acquired in triplicate by 2 raters in 9 postmenopausal women with osteoporosis of the spine estimated using generalizability theory (G-Theory) and classical test theory (CTT).

Measures of spine curvature	Inter-trial reliability		Inter-rater reliability	
	G-theory	CTT	G-theory	CTT
Cervicothoracic angle				
Reliability coefficient	0.960	0.960	0.566	0.601
SEM (degrees)	2.281	2.040	7.505	7.091
Thoracolumbar angle				
Reliability coefficient	0.958	0.964	0.726	0.722
SEM (degrees)	3.090	2.703	7.868	7.786
Lumbosacral angle				
Reliability coefficient	0.942	0.946	0.637	0.630
SEM (degrees)	2.498	2.367	6.223	6.213
Kyphotic index				
Reliability coefficient	0.956	0.959	0.921	0.920
SEM	1.097	1.040	1.474	1.461
Lordotic index				
Reliability coefficient	0.840	0.837	0.746	0.768
SEM	1.427	1.390	1.794	1.701

SEM: standard error of the measurement provides an estimate of absolute reliability and is expressed in the same units as the measure.

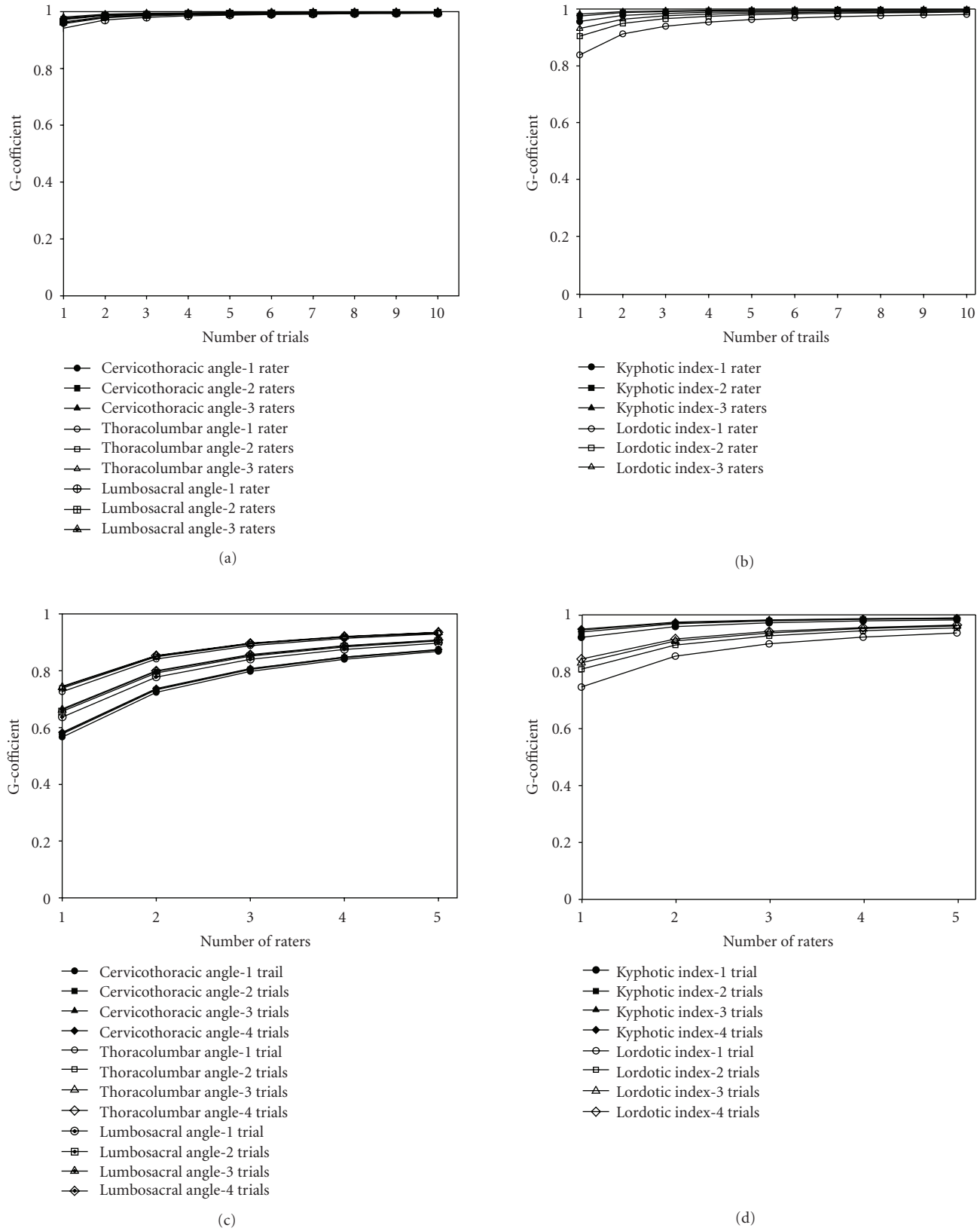

Figure 1: The results of the design study for optimizing inter-trial reliability are illustrated in which the influence of having different numbers of raters is shown as a function of the number of trials for (a) spine curvature angles (degrees) measured using the digital inclinometer and (b) kyphotic index and lordotic index measured using the flexicurve ruler. The results of the design study for optimizing inter-rater reliability are illustrated in which the influence of performing different numbers of trials is shown as a function of raters for (c) spine curvature angles (degrees) measured using the digital inclinometer, and (d) kyphotic index and lordotic index measured using the flexicurve ruler.

5. Discussion

This study aimed to illustrate the application of the tools of G-theory to establish a measurement protocol with optimal inter-trial and inter-rater reliability for assessing spine curvatures in postmenopausal women with osteoporosis of the spine. Estimates of inter-trial and inter-rater reliability of spine curvature measures acquired using the digital inclinometer and flexicurve ruler were similar whether using G-theory or CTT approaches. G-Theory provides an advantage in utilizing even small datasets to explore the effect of changing aspects of the study design (e.g., number of raters and number of trials) in order to identify the optimal measurement protocol for a particular clinical or research setting.

Reliability of outcome measures needs to be established for each specific clinical environment or research laboratory. In our example, all measures of spine curvature had acceptable reliability (high reliability coefficients and low SEM) when performed by the same rater in triplicate (Table 3). No literature was found describing the reliability of measures of spine curvatures in postmenopausal women with osteoporosis of the spine using the digital inclinometer. However, KI inter-trial reliability (0.96) and inter-rater reliability (0.92) were comparable to or exceeded that reported by investigators using CTT (Lundon et al. [3]: $0.86 \leq \text{ICC} \leq 0.97$; Arnold et al [4]: $0.86 \leq \text{ICC} \leq 0.91$). These findings suggest that brief training was adequate for acquiring reliable measures of KI. For all the measures, it would be preferable to have the same rater perform the measurements in women with osteoporosis of the spine whether being followed in the clinic or enrolled in a longitudinal research study.

Inter-rater reliability for LI measures was adequate given the G-coefficient of 0.75 in combination with a low SEM (1.72). However, inter-rater reliability of spine curvature measures acquired using the inclinometer was not adequate with G-coefficients varying from 0.57 to 0.73 and SEM varying from 6.22 to 7.87 degrees. The use of D-studies provided an efficient way to optimize the measurement process. We determined that inter-rater reliability could be improved satisfactorily for the TL angle and LS angle by having 5 raters acquire the measures 4 times. Scenarios for optimizing inter-rater reliability of CT angle fell outside the realm of clinical feasibility. We did not have to conduct different studies to determine whether greater gain in reliability would be achieved by increasing number of raters or increasing the number of assessments. We were able to acquire this information based on measures obtained in only 9 women representative of our target study population.

A limitation of this study may be the inclusion of assessors with varying levels of clinical experience. Neither assessor had used the flexicurve ruler before, however, the physiotherapist had over 20 years of experience performing physical assessments in general clinical practice. By building the different experience levels into the study design, we could illustrate nonzero sources of variance. However, the mean spine curvature measures acquired by each rater varied considerably, particularly when using the digital inclinometer, and this study was not designed to determine the accuracy of the measures. It would be interesting to determine the results following more extensive training of novice raters, inclusion of an expert rater, and verification of landmarks identified by each rater. Nonetheless, these results provide estimates of reliability that can be generalized to assessors with minimal levels of experience assessing posture and demonstrate that when the same rater measures spine curvatures, the measures are consistent.

6. Conclusions

We intend the results of this study to be used at the discretion of clinicians and investigators who are using measures of spine curvatures obtained using the flexicurve ruler or digital inclinometer in the clinical assessment of individuals with osteoporosis. Furthermore, this approach may be replicated to identify other measurement protocols that optimize reliability. Ultimately a suitable compromise between a feasible measurement protocol and acceptable reliability for each particular clinical or research setting must be identified. G-theory provides an alternative to CTT that enables efficient identification of an optimal measurement protocol based on data collected in a reliability study having a single study design.

Acknowledgments

The authors thank the participants who volunteered for their study and Leslie Beaumont for her assistance in recording the spine curvature measures for each rater. This study was funded in part by the Natural Science and Engineering Research Council of Canada (NSERC)—Discovery Grant (NJM).

References

[1] J. C. Nunnally, *Psychometric Theory*, McGraw-Hill, Toronto, Canada, 1978.

[2] S. Messick, "Validity," in *Educational Measurement*, R. L. Linn, Ed., ORYZ Press, Phoenix, Ariz, USA, 3rd edition, 1993.

[3] K. M. A. Lundon, A. M. W. Y. Li, and S. Bibershtein, "Interrater and intrarater reliability in the measurement of kyphosis in postmenopausal women with osteoporosis," *Spine*, vol. 23, no. 18, pp. 1978–1985, 1998.

[4] C. M. Arnold, B. Beatty, E. L. Harrison, and W. Olszynski, "The reliability of five clinical postural alignment measures for women with osteoporosis," *Physiotherapy Canada*, vol. 52, pp. 286–294, 2000.

[5] M. R. Hinman, "Interrater reliability of flexicurve postural measures among novice users," *Journal of Back and Musculoskeletal Rehabilitation*, vol. 17, no. 1, pp. 33–36, 2003.

[6] L. J. Cronbach, R. Nageswari, and G. C. Gleser, "Theory of generalizability: a liberation of reliability theory," *The British Journal of Statistical Psychology*, vol. 16, pp. 137–163, 1963.

[7] R. L. Brennan, *Statistics for Social Science and Public Policy: Generalizability Theory*, Springer, New York, NY, USA, 2001.

[8] R. J. Shavelson, N. M. Webb, and G. L. Rowley, "Generalizability theory," *American Psychologist*, vol. 44, no. 6, pp. 922–932, 1989.

[9] E. Itoi, "Roentgenographic analysis of posture in spinal osteoporotics," *Spine*, vol. 16, no. 7, pp. 750–756, 1991.

[10] F. J. Bonner Jr., M. Sinaki, M. Grabois et al., "Health professional's guide to rehabilitation of the patient with osteoporosis," *Osteoporosis International*, vol. 14, supplement 2, pp. S1–S22, 2003.

[11] C. Lindsey and N. Bookstein, "Kypholordosis Measurement Using a Flexible Curve (Instructional CD)," American Physical Therapy Association Section on Geriatrics, 2007.

[12] A. M. Briggs, J. H. Van Dieën, T. V. Wrigley et al., "Thoracic kyphosis affects spinal loads and trunk muscle force," *Physical Therapy*, vol. 87, no. 5, pp. 595–607, 2007.

[13] R. K. Chow and J. E. Harrison, "Relationship of kyphosis to physical fitness and bone mass on post-menopausal women," *American Journal of Physical Medicine*, vol. 66, no. 5, pp. 219–227, 1987.

[14] D. M. Kado, M. H. Huang, E. Barrett-Connor, and G. A. Greendale, "Hyperkyphotic posture and poor physical functional ability in older community-dwelling men and women: the Rancho Bernardo Study," *Journals of Gerontology A*, vol. 60, no. 5, pp. 633–637, 2005.

[15] H. D. Saunders, "Saunder's digital inclinometer: user's guide," United States, Empi Therapy Solutions, 2008.

[16] R. Bloch, "G_String_III [computer program]. Version 5.4.2 for Windows. [Hamilton, ON:] Accompanied by: 1 user manual (pdf)," 2010, http://www.fhs.mcmaster.ca/perd/download/g_string