The Comprehensive Osteoarthritis Test: a Simple Index for Measurement of Treatment Effects in Clinical Trials

LYNDON O. BROOKS, MARGARET I. ROLFE, PHILLIP A. CHERAS, and STEPHEN P. MYERS

ABSTRACT. ***Objective***. To assess the measurement properties of a simple index of symptom severity in osteoarthritis (OA) of the hips and knees.

Methods. Both the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) and the proposed new Comprehensive Osteoarthritis Test (COAT) instrument were completed weekly by 125 subjects in the context of a randomized, 12-week, 3 parallel-arm clinical trial. The reliabilities of the various scales were assessed on a weekly basis by use of Cronbach's alpha coefficients. The validity of the COAT total scale was assessed by correlation with the WOMAC total scale on a weekly basis with correlation coefficients, and in terms of the correlations between subject-level intercepts and slopes over time. The relative responsiveness of the WOMAC and COAT total scales was assessed using a multilevel (longitudinal) multivariate (WOMAC, COAT) linear model.

Results. The WOMAC and COAT total scales were highly reliable (mean over weeks: WOMAC alpha = 0.98; COAT alpha = 0.97). The correlations between the WOMAC and COAT scales were very high (mean over weeks = 0.92; subject-level intercepts = 0.91, slopes = 0.88). The COAT total scale was significantly more responsive than the WOMAC total scale in the active treatment (34.8% improvement vs 26.8%; p = 0.002).

Conclusion. The COAT total scale is simple to administer, reliable, valid, and responsive to treatment effects. (J Rheumatol 2004;31:1180–6)

Key Indexing Terms:
COMPREHENSIVE OSTEOARTHRITIS TEST SYMPTOM SCALE OSTEOARTHRITIS
VALIDITY RELIABILITY RESPONSIVENESS

The Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC)[1] has been widely employed in osteoarthritis (OA) clinical trials since publication of validation results in 1988[1], and might be considered to be the gold standard among disease specific, self-report instruments for measurement of the symptoms of OA of the hips and knees. The WOMAC consists of 24 items [measured on 5-point Likert or alternative 100 mm visual analog (VAS) response scales] chosen as characteristic of the more physical expressions of the disorder (social and emotional subscales were abandoned following unsatisfactory performance) and found to contribute to formation of reliable subscales for pain (5 items), stiffness (2 items), and physical dysfunction (17 items). Subscale scores are calculated by summing over the respective item sets and the total index by summing over the entire item set (although some items may be removed as irrelevant to individual respondents). Scores may be expressed in terms of the original item response scale by dividing these aggregates by the appropriate numbers of items. This way of calculating the total score (total WOMAC) means that it is most strongly associated with the disability subscale and least strongly associated with the stiffness subscale according to the relative numbers of items of which they are composed.

Scores constructed as composites of several items have the advantages that they may be both more reliable and more clearly valid than single item measures. In respect of reliability, summing or averaging "damps down" extraneous sources of variation that may be associated with each of the items considered separately. In respect of validity, the construct measured by a scale may be more immediately evident if it includes a sufficient set of items to compose a relatively complete description of an identifiable aspect of the typical manifestation of the disease. However, should it be possible to construct simpler measures with suitable levels of reliability, validity, and responsiveness, they may have the advantages of ease of administration and decreased respondent burden with repetitive administration.

By helping to alleviate respondent burden, or increasing respondent satisfaction with the response task, simpler

From the Australian Centre for Complementary Medicine Education and Research, University of Queensland, Brisbane; and Southern Cross University, Lismore, Australia.

L.O. Brooks, PhD, Research Methodologist, Australian Centre for Complementary Medicine Education and Research and Southern Cross University; M.I. Rolfe, MStat, Statistician, Southern Cross University; P.A. Cheras, PhD, Deputy Director, Australian Centre for Complementary Medicine Education and Research and University of Queensland; S.P. Myers, PhD, Professor and Director, Australian Centre for Complementary Medicine Education and Research, University of Queensland and Southern Cross University.

Address reprints requests to Dr. L. Brooks, Graduate Research College, Southern Cross University, PO Box 157, East Lismore, Australia. E-mail: lbrooks@scu.edu.au

Submitted March 19, 2003; revision accepted December 19, 2003.

measures may help to alleviate the sample attrition problem that is a feature of longitudinal research. An often neglected aspect of respondent satisfaction with measurement instruments is the extent to which patients feel that the items and the response scale allow them to express how their condition affects them personally. While it is recognized that the items composing WOMAC together constitute a valid description of the typical expression of OA, that expression may vary to some extent with the individual case and lifestyle. A number of items in the physical dysfunction subscale in particular do not apply at all to some patients (those who shower rather than bath, for example) or apply in varying degrees depending upon lifestyle and circumstances. Effects of this may be to foster a sense of irrelevancy with respect to some items or frustration that the items do not include personally relevant concerns. An important additional effect may be that, because there may be lower baseline scores and less change measured with treatment on less personally relevant items, therapeutic changes on the total scale may underestimate effects of more personally relevant expressions of the disease.

While it is important at some stage in the development of a valid scale to identify a set of items that define the illness and distinguish it from other sources of distress or dysfunction, the validity of a simpler instrument that lacks this source of validation may be assessed in part by its correspondence with a scale accepted as valid. Our study relies upon the validity of WOMAC, whose developers carefully selected items according to clinical and empirical criteria[1] for validation of the Comprehensive Osteoarthritis Test (COAT).

Well-designed longitudinal research may often allow relatively unambiguous attribution of causality and estimation of non-linear response profiles, such as when a period of acute response is followed by a slower rate of improvement that levels off at some point. Relatively recent developments in statistical modeling techniques have made the analysis of longitudinal data more tractable than previously. For example, multilevel models with maximum likelihood (or restricted maximum likelihood) estimation provide for reliable estimates of the effects of treatment over time in the context of missing data[2-4]; and fitting random effects may often remove residual dependency. A model of this type is reported subsequently.

Against this background, the objective was to develop an index of OA symptom severity specifically for use in longitudinal clinical trials. The index was to be short, simple to respond to and score, reliable, valid, and responsive to treatment effects.

WOMAC employs 100 mm VAS as an alternative to 5-point Likert scales. The WOMAC VAS have been analyzed here and the new instrument based on them. There is evidence from psychophysics and psychometrics that visual analog response scales map linearly onto stimulus intensities[5], justifying the assumption of interval measurement for VAS scores. Indeed, subjects may provide ratio level measurement of the "magnitudes" of subjective phenomena on VAS under appropriate circumstances[6]. Moreover, VAS may be more reliable than verbal (e.g., 5-point Likert) or numerical (e.g., 0 to 10) rating scales[6], and be more sensitive to change[7], thus requiring smaller sample sizes to achieve the same statistical power. However, there is some evidence that VAS are less reliable in older and less well-educated populations[8].

COAT was constructed on the assumptions that pain, stiffness, and physical dysfunction are the 3 main components of OA symptoms, and that patients are able to reliably estimate the effect of OA in these terms: i.e., at this level of analysis, over and above their ability to estimate the effects at the more basic, context-specific level of the WOMAC items. An even more global judgment was also included: to estimate the overall severity of OA symptoms. While this may function in some cases simply as a summary of the 3 main components or serve to re-weight them to some extent, in other cases it may allow for respondent inclusion of effects that they attribute to OA that are not completely captured in terms of pain, stiffness, and physical dysfunction. Because we do not further describe the performance of the 3-item version of the COAT total scale, it is appropriate to mention here that the results were highly similar between both the 3- and 4-item versions of the scale but typically slightly in favor of the 4-item version.

MATERIALS AND METHODS

The COAT scale consists of four 100 mm VAS, one each for pain, stiffness, difficulties with physical activities, and the overall effect of OA. The scales were introduced as follows: Osteoarthritis typically causes joint pain and stiffness, and makes many physical activities difficult. Osteoarthritis may also affect you in other ways. We need to know how much osteoarthritis has affected you over the PAST WEEK in terms of joint pain, joint stiffness, difficulty in performing physical activities, and the overall severity of your symptoms (all things considered). Please answer the questions in respect of your nominated joint. Please mark the scales below to show the severity of your osteoarthritis symptoms over the past week.

The 4 scales were presented labelled as "joint pain," "stiffness," "difficulties with physical activities", and "overall symptoms", and were anchored at their ends with "none" and "extreme."

Scores on the items were measured in millimetres, and the total score was calculated as the (unit-weighted) mean of the 4 item scores.

Interviews conducted with patients at induction included questions about problems they had in responding to the WOMAC and COAT instruments. Patients with OA at multiple sites were told to think about the joint that caused them the most problems. None reported difficulty in answering questions in respect of their nominated joint, including the COAT item about their overall symptoms. Accordingly, the overall symptoms item is to be interpreted as a summary of symptoms in respect of the nominated joint rather than as patient global assessment measure.

The data were generated in the course of a double-blind, randomized, 3 parallel-arm trial extending over 15 weeks (4 wks washout followed by 11 wks of treatment). The study compared responses to 3 treatments: a placebo, a marine extract, and combination of marine and herbal extracts. A total of 125 subjects (71 male and 54 female) were enrolled, of whom 108 completed the trial.

Subjects were recruited through newspaper advertisements for a randomized placebo controlled clinical trial of an oral complementary medicine for the treatment of hip and knee OA. Subjects were required to have radiographically proven OA, be in good general health, and be taking no complementary or pharmaceutical treatment for their OA, or be willing to cease for the period of the trial. At screening, 28.2% reported the use of complementary medicine, 36.4% were using nonsteroidal antiinflammatory drugs, and 58.2% reported a family history of arthritis. Of the subjects in the trial, 80% had OA of the knee, 43.6% OA of the hip, and 23.6% OA of both hip and knee.

The WOMAC and COAT scales were administered at baseline and weekly for 11 weeks subsequently. Some subjects missed measurement occasions and others "dropped out" during the treatment period.

Statistical analysis of the data is presented in 3 parts: reliability, validity, and responsiveness.

The reliabilities of the various scales (WOMAC pain, stiffness and dysfunction, and total, and COAT total) were assessed using Cronbach's alpha coefficients on a week-by-week basis. Alpha, a measure based on the mean inter-item correlation, is an appropriate measure of internal consistency reliability because all scales are unit-weighted.

The validities of the COAT items and total scale were assessed by correlation with WOMAC sub- and total scales. In the first instance, the analyses were conducted on a week-by-week basis and included the correlations among the items and (sub)scales both within and between the WOMAC and COAT sets. While these analyses were cross-sectional, an additional aspect of the relationship between the WOMAC and COAT scales (here only the total scales were considered) is the correlation between the intercepts and response slopes over time at the subject level. The hierarchical linear model subsequently described estimated a linear regression of the WOMAC and COAT scales over time for each subject. The estimated mean intercepts and slopes for the 2 scales were the fixed effects around which the individual intercepts and slopes were estimated as random effects. For both the random (subject level) intercepts and slopes, the correlation between the WOMAC and COAT profiles was calculated as the covariance between the 2 divided by the square root of the product of their variances. This model and procedure are described by Goldstein[2] and others[9,10].

Two models were fit to assess the relative responsiveness of the WOMAC and COAT total scales, the first of which generated the statistics used to calculate the correlations between the WOMAC and COAT subject-level intercepts and slopes. The data analyzed in this model were from the more effective treatment only. A model analyzing the data from all 3 treatments was considerably more complex without further contributing to the present argument.

The model employed may be described as a multivariate (WOMAC, COAT), longitudinal (time = 0 to 11), multilevel (subjects, observations within subjects), linear response (intercept, time) model[3,9,10]. Time was fit as varying randomly as change over time was highly significantly variable among subjects on both sets of scores; i.e., the model fitted random intercepts (subjects, observations within subjects) and slopes (within subjects). Here, the linear regression of the WOMAC and COAT scales over time was estimated for each subject: i.e., the estimated mean intercepts and slopes for the 2 scales were the fixed effects around which the individual intercepts and slopes were estimated as random effects.

A further multilevel (or hierarchical) model was fit to the data from the more effective treatment only to directly compare the response slopes for WOMAC and COAT. Both sets of scores were treated as one vector of response (per subject) with the WOMAC and COAT sets distinguished by a dummy variable "measure." The fixed effects fitted were the intercept, measure, (linear) time, and measure by time. The important parameter from the present perspective was the measure by time interaction, which measures the difference in mean response slopes between the WOMAC and COAT scores.

RESULTS

Reliability. Cronbach's alpha coefficients were calculated for the WOMAC pain, stiffness, and dysfunction, and the WOMAC and COAT total scales at each of the 12 occasions of measurement. Only 21 of the 24 items composing the WOMAC total scale were used in these analyses as between 30 and 60 subjects on any occasion declined to respond to at least one of the 3 dysfunction scale items "bath," "heavy," or "shop". The coefficients are summarized in terms of their minima, maxima, means, and standard deviations in Table 1. All scales had consistently high reliabilities (minimum alpha = 0.86 for the WOMAC stiffness scale at baseline). The mean alpha over the 12 occasions was 0.98 for the WOMAC total scale and 0.97 for the COAT total scale.

Validity. Correlation coefficients were calculated to estimate the relations among the WOMAC scales, among the COAT items and total scale, and between corresponding WOMAC and COAT scores. The correlations within sets are summarized in the text below, and those between sets are reported in Table 2. In each case, the coefficients are summarized in terms of their minima, maxima, means, and standard deviations over occasions.

The correlations among the WOMAC scales and among the COAT items and total scale (within sets) were very substantial (WOMAC minimum = 0.80 for pain by stiffness, WOMAC maximum = 0.99 for dysfunction by total, COAT minimum = 0.86 for pain by difficulties, COAT maximum = 0.98 for overall by total). In general, the within sets correlations were very high for subscales presumed to measure empirically separable aspects of OA symptoms. In particular, it is doubtful whether the WOMAC dysfunction scale is distinguishable from the WOMAC total scale (as might be expected when 14 or 17 of the 21 or 24 items in the total scale measure contexts of dysfunction). Although there may be differential response rates among the pain, stiffness, and dysfunction/difficulties subscales for some classes of patients and some treatments, such effects were minimal in these data. Under these circumstances the subscales function substantially as proxies for each other, and it may be an error to conclude that a treatment is selectively effective for pain or stiffness or dysfunction when one but not others is found to measure a significant effect. It is prudent to analyze only the total scales for this reason.

The WOMAC-COAT correlations (between sets) for each of the scales were substantial (minimum r = 0.84 for stiffness). Most important, because (as argued above) it may be imprudent to separately analyze subscale scores, the mean correlation between the WOMAC and COAT total scales was 0.92. This is good evidence that, to the extent that the WOMAC total scale is considered to be valid, the COAT scale may be considered to be valid also.

While the relationship between the total scales is of primary importance, nonetheless for each of the WOMAC scales, the correlations were higher with the corresponding COAT scale in all cases except for dysfunction/difficulties. In this case, the WOMAC dysfunction scale correlated

Table 1. Cronbach's alpha coefficients: minima, maxima, means, and standard deviations over the 12 occasions of measurement for the WOMAC pain, stiffness, dysfunction, and the WOMAC and COAT total scales.

Scale (No. of Items)	Minimum	Maximum	Mean	SD
Pain	0.875	0.950	0.937	0.021
Stiffness	0.861	0.969	0.941	0.034
Dysfunction (17)[a]	0.970	0.984	0.979	0.003
Dysfunction (14)[a]	0.957	0.982	0.977	0.007
Total (24)[a]	0.974	0.987	0.983	0.003
Total (21)[a]	0.965	0.987	0.983	0.006
Total COAT	0.946	0.978	0.969	0.009

[a] Between 30 and 60 subjects declined to respond to at least one of the 3 dysfunction scale items, bath, heavy, or shop, depending upon occasion. The dysfunction 17 item scale and total 17 item scale results include all items but, due to missing data, were based on much smaller samples than the dysfunction 14 item scale and total 14 item scale results, which exclude the bath, heavy and shop items.

Table 2. Correlations between the WOMAC and COAT scales: minima, maxima, means, and standard deviations by occasion.

		COAT Scale				
	Statistic	Pain	Stiffness	Difficulties	Overall	Total
WOMAC Scale						
Pain	Minimum	0.742	0.643	0.624	0.690	0.728
	Maximum	0.912	0.828	0.873	0.908	0.908
	Mean	**0.853**	0.773	0.794	0.839	0.850
	SD	0.052	0.052	0.068	0.061	0.051
Stiffness	Minimum	0.606	0.680	0.648	0.635	0.692
	Maximum	0.875	0.903	0.820	0.883	0.884
	Mean	0.782	**0.844**	0.756	0.805	0.830
	SD	0.091	0.063	0.055	0.073	0.062
Dysfunction	Minimum	0.706	0.778	0.802	0.806	0.832
	Maximum	0.902	0.885	0.911	0.934	0.932
	Mean	0.852	0.861	**0.890**	0.896	0.913
	SD	0.055	0.029	0.030	0.034	0.027
Total	Minimum	0.742	0.797	0.799	0.816	0.850
	Maximum	0.919	0.887	0.913	0.941	0.943
	Mean	0.871	0.868	0.884	0.902	**0.920**
	SD	0.051	0.024	0.032	0.035	0.026

Values in bold face indicate mean correlation between corresponding WOMAC and COAT subscales over 12 occasions.

slightly more highly with the COAT total than with the COAT difficulties scale. These results are further evidence that the WOMAC and COAT instruments measure the same aspects of OA symptoms in very similar ways.

A further aspect of the relationship between the WOMAC and COAT scales (here only the total scales are considered) are the correlations between the intercepts and response slopes at the subject level. Calculated as described in Materials and Methods, the subject level intercepts correlated 0.91 and the slopes correlated 0.88 between the total WOMAC and COAT scales.

Together, the strong correlations between the WOMAC and COAT scales at each occasion and between WOMAC and COAT response variables over time represent good evidence of the validity of the COAT scale, conditional upon the validity of the WOMAC scale.

Responsiveness. The relative responsiveness to treatment effects of WOMAC and COAT total scales was assessed, in the first instance, by fitting a model to the 2 sets of scores simultaneously. The parameter estimates and their standard errors (SE) and null hypothesis probabilities are reported in Table 3. The model-estimated linear profiles on the 2 sets of scores (WOMAC, COAT) are plotted in Figure 1.

The parameter estimate for the focal time effect was 54% larger for COAT than WOMAC. Although its SE was also larger (14%), the z value was considerably larger for COAT than WOMAC (5.47, 4.03). Mean scores decreased by 18.5 (± 6.2 95% CI; p = 0.013) from baseline (53.0) to week 11

Table 3. Multilevel random intercept and random linear slope model for the more effective treatment on 2 sets of scores (WOMAC, COAT): parameter estimates, standard errors, and null hypothesis probabilities.

Measure		WOMAC			COAT	
Effect	Estimate	SE	p	Estimate	SE	p
Fixed						
Intercept	44.656	3.166	0.000	53.023	3.292	0.000
Time	–1.087	0.270	0.000	–1.678	0.307	0.000
Random†						
Subject:						
Intercept (I)	390.724	90.721	0.000	405.640	98.101	0.000
Slope (S)	2.414	0.652	0.000	2.827	0.843	0.000
I by S cov.	–3.533	5.556	0.262	–6.279	6.708	0.175
Occasion:						
I	62.936	4.539	0.000	120.576	8.732	0.000
		WOMAC by COAT				
Subject:						
I cov.	362.083	89.729	0.000	I correlation‡		0.91
S cov.	2.307	0.692	0.000	S correlation‡		0.88

† Variances and covariances. ‡ Correlation = $\frac{\text{covariance}_{ab}}{\sqrt{\text{variance}_a \times \text{variance}_b}}$

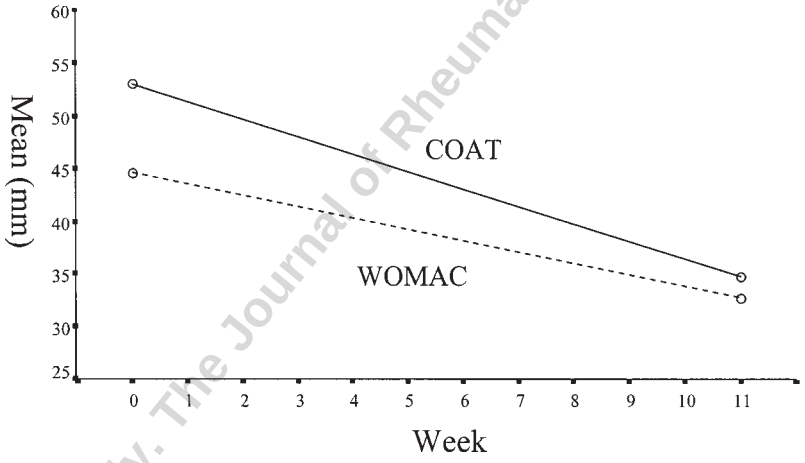

Figure 1. Linear estimated mean score by week (0–11) for the active treatment by measure (WOMAC, COAT).

or 34.8% for the COAT measure, and 12.0 (± 5.2 95% CI; p = 0.022) or 26.8% from baseline (44.7) to week 11 for the WOMAC measure. Fitting a similar model to the data from another treatment showed the overall changes to be both relatively small (–1.3 for COAT, +1.6 for WOMAC) and non-significant (COAT p = 0.691; WOMAC p = 0.540). Consequently, while COAT appears to be more responsive than WOMAC to effective treatments, it appears to reliably estimate null effects also.

It is a feature of our data that mean baseline scores were observed higher on the COAT than WOMAC total scale. That the intercept by slope covariances for both the WOMAC and COAT scales were small relative to their SE and non-significant (Table 3: covariance; SE, p: WOMAC –3.53, 5.56, 0.262; COAT –6.28, 6.71, 0.175) indicates that the measured responses to treatment were not statistically reliable functions of baseline measured severities. Nonetheless, it is likely that the higher baseline scores on COAT than WOMAC and its superior responsiveness are related through a common process, such as implicit selection of more personally relevant expressions of the disease.

A further model was fitted to the data from the more effective treatment to directly compare the response slopes for WOMAC and COAT. The measure (WOMAC, COAT) by time interaction effect was highly significant in favor of the COAT scale (estimate = –0.55 difference per week, SE

= 0.19, p = 0.002), indicating that, in these data, the COAT scale was significantly more responsive than the WOMAC scale.

The effect sizes and standardized response means were 0.63 and 7.73, respectively, for the WOMAC total scale and 0.82 and 10.86, respectively, for the COAT total scale. These statistics are each measures of mean response rates relative to variation (baseline variation for the effect size and response slope variation for the standardized response mean). As such they clearly indicate more powerful analyses on COAT than WOMAC total scores.

Although, as was previously argued, the (within sets) weekly correlations among the sets of WOMAC and COAT subscales were so high as to make their separate analysis questionable, it is possible that the subscales (within sets) were differentially responsive to treatment. Figure 2 shows the subscale means plotted by week for the WOMAC and COAT sets. Descriptively, the changes over time were similar for each of the subscales in each of the WOMAC and COAT sets. Nonetheless, a tendency may be observed for the physical dysfunction/difficulties subscales to be least responsive, and the stiffness subscales to be most responsive, to treatment. It may also be of interest to observe that the response on COAT overall subscale was typical of the responses on the other COAT subscales, suggesting that it may function to summarize them rather than to enhance their collective responsiveness.

DISCUSSION

Given comparable reliability and validity for the WOMAC and COAT instruments, COAT has the advantages of ease of administration, reduced respondent burden, and superior responsiveness to therapeutic change.

The relative ease of administration of the COAT instrument enhances its potential for more frequent measurement and for responding from home in the form of self-completed

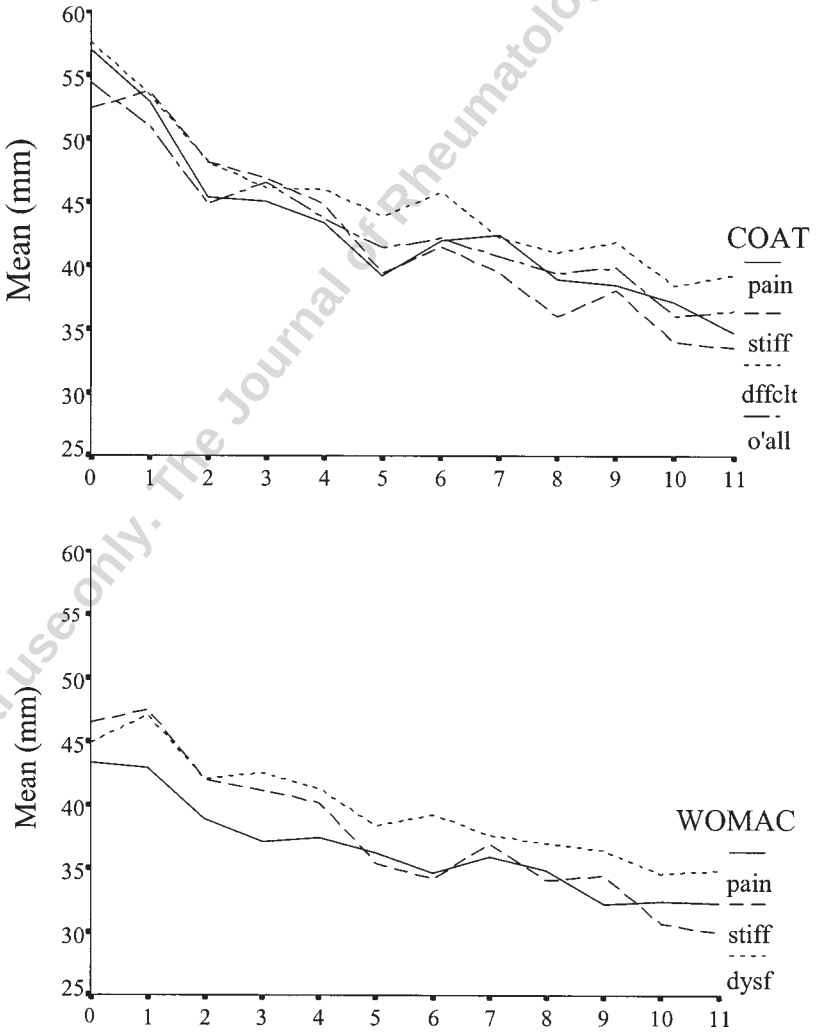

Figure 2. Weekly observed means of the COAT and WOMAC sets of subscales for the more effective treatment.

diary entries, onscreen via the internet or, with further development, by telephone. While the increased responsiveness of COAT over WOMAC increases the statistical power of trial designs and so requires smaller samples with associated lower costs, the ease of administration of COAT and its potential for remote measurement make it possible to conduct larger trials with more frequent measurement in the community, as well as/instead of in the clinic.

In turn, more frequent measurement on larger samples has the potential to make trial results more informative, especially in respect of identifying the slopes and shapes of response profiles over time. Is there a typical form of the response profile? If so, after what time do treatment effects plateau? If not, are there distinctly different profiles of response and are profile differences between classes of responders predictable from available variables?

Although we are satisfied as to the reliability and validity of COAT, researchers who hesitate to employ it until further data on its performance are available may consider using both instruments. One strategy may be to employ both instruments at critical times in the trial, such as at baseline and completion or when patients visit the clinic, and to employ COAT on intermediate measurement occasions when measurement may be remote. This strategy would provide alternative measures of primary outcomes, a basis for assessment of the contextual validity of COAT against WOMAC, and a sequence of scores for modeling response profiles. Alternatively, 100 mm VAS are commonly used and considered valid as measures of pain. COAT simply extends the usage of 100 mm VAS to the other 2 principal components of OA symptoms, stiffness, and difficulties with physical activities. It is the considerable covariation among these components in OA patients and their strong tendency to change together with treatment that identifies the composite score as an OA-specific measure.

A comment on the COAT overall subscale is in order. It is unlikely that this subscale measures a distinct component of OA symptoms to any great extent, appearing more to summarize the primary pain, stiffness, and difficulties components. Its performance in this respect may depend upon its being included with the other subscales on the patient response form. Consequently, what the overall subscale might measure (or perhaps constitute a valid measure of) if used independently is presently unclear, and is not — without further data — recommended as an independent patient global assessment scale. It does not appear to be more responsive than the other subscales and its utility is principally to contribute to the reliability of the total scale.

It was suggested that the COAT total scale may be more responsive than the WOMAC total scale in part because it does not include items at the level of specificity of those included among the WOMAC set that may be of varying relevance to patients, with items perceived as less relevant to a patient showing less responsiveness. The initiative to assess symptom severity at the global pain, stiffness, difficulties, and overall levels was based on the assumption that patients would intuitively form a response in terms of expressions of the disease as they experience it and as are relevant to them. Future research will assess patient preference and examine the role of perceived item relevance as a factor in the responsiveness of the scales.

Given the global level at which the 3 components of OA symptoms (pain, stiffness, difficulties with physical activities, overall) are measured by COAT, it is arguable that the index may provide reliable and valid measurement of the symptoms of OA of the hands as well as of the hips and knees. This is not the case with the WOMAC index in which the items are more specific to hip and knee function. However, the authors of WOMAC have now published AUSCAN for hand arthritis[11]. Future research might assess the potential for use of the COAT scale for hand arthritis by comparison with the AUSCAN. Should COAT perform well in such a comparison it would have the advantage of providing a single measure of symptoms for clinical trials of OA of the hips, knees, and hands.

In sum, the COAT instrument is short, simple to respond to and score, and is reliable, valid, and responsive to the effects of treatment in OA. It may be especially useful in clinical trials in which frequent, remote, or extended measurement over time is planned.

REFERENCES

1. Bellamy N, Buchannan WW, Goldsmith CH, Campbell J, Stitt LW. Validation study of WOMAC: A health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. J Rheumatol 1988;15:1833-40.
2. Goldstein H. Multilevel statistical models. 2nd Ed. London: Edward Arnold; 1995.
3. Snijders T, Bosker R. Multilevel analysis: An introduction to basic and advanced multilevel modeling. London: Sage; 1999.
4. Leyland AH, Goldstein H, editors. Multilevel modelling of health statistics. Chichester: John Wiley and Sons; 2001.
5. McDowell I, Newell C. Measuring health: A guide to rating scales and questionnaires. NY; Oxford: Oxford University Press; 1996.
6. Price DD, McGrath PA, Rafii A, et al. The validation of visual analogue scales as ratio scale measures for chronic and experimental pain. Pain 1983;17:45-56.
7. Sriwatanakul K, Kelvie W, Lasagna L, et al. Studies with different types of visual analogue scales for measurement of pain. Clin Pharmacol Ther 1983;34:234-9.
8. Ferraz MB, Quaresma MR, Aquino LRL, et al. Reliability of pain scales in the assessment of literate and illiterate patients with rheumatoid arthritis. J Rheumatol 1990;17:1022-4.
9. Rasbash J, Browne W, Goldstein H, et al. A user's guide to MLwiN. London: Multilevel models project. London: Institute of Education, University of London; 1999.
10. MacCallum R, Kim C. Modeling multivariate change. In: Little TD, Schnabel KU, Baumert J, editors. Modeling longitudinal and multilevel data: Practical issues, applied approaches and specific examples. Mahwah, NJ; Lawrence Erlbaum Associates; 2000.
11. Bellamy N, Campbell J, Haroui B, et al. Clinimetric properties of the AUSCAN Osteoarthritis Hand Index: an evaluation of reliability, validity and responsiveness. Osteoarthritis Cart 2002;10:863-9.