OsteoArthritis and Cartilage (2003) **11,** 499–507
© 2003 OsteoArthritis Research Society International. Published by Elsevier Science Ltd. All rights reserved.
doi:10.1016/S1063-4584(03)00082-7

Osteoarthritis and Cartilage

International Cartilage Repair Society

OARSI OSTEOARTHRITIS RESEARCH SOCIETY INTERNATIONAL

Response relationship of VAS and Likert scales in osteoarthritis efficacy measurement

J. A. Bolognese M.Stat. Senior Director Scientific Staff†*, T. J. Schnitzer M.D., Ph.D. Professor of Medicine‡ and E. W. Ehrich M.D. Vice President§

†*Merck Research Labs, RY34-304, P.O. Box 2000, Rahway, NJ 07065, USA*

‡*Northwestern University, Feinberg School of Medicine, 710 N. Lake Shore Drive, Room 1020, Chicago, IL 6061, USA*

§*Medical Affairs, Alkermes Inc., 64 Sidney Street, Cambridge, MA 02139, USA*

Summary

Objective/Background: Efficacy in osteoarthritis (OA) is principally measured using subjective visual analogue (VAS) and/or Likert scale responses. The relationship between these two scales and their relative precision in discriminating active from placebo treatment in OA patients was determined.

Design/Methods: Patient overall pain assessment, and patient and investigator global assessments were each measured on a 100 mm VAS and on a 0 to 4 point Likert scale in a 6-week OA study of rofecoxib vs placebo. The relationship between the VAS and Likert responses was examined graphically and via summary statistics. Analysis of variance was used to assess consistency of the VAS/Likert relationship over time and across the different endpoints. Precision was compared using effect size, and normality of VAS scale of measurement was assessed using the Shapiro–Wilk test.

Results: Mean VAS scores and changes from baseline at individual time points were generally highly correlated with corresponding Likert responses (r-values generally approximately 0.7–0.8). The magnitude of VAS values and changes varied depending on endpoint, on the associated magnitude of increment of Likert score, and on the Likert baseline value (i.e., where on the Likert scale the change was occurring). Precision of VAS and Likert responses to detect difference between treatments was generally similar with effect sizes approximately 1. Normality and homogeneity of variance of VAS scores was most closely approximated by actual changes in comparison to percent change or log-transformed measures.

Conclusions: VAS and Likert responses are highly correlated and yield similar precision for discriminating treatments in OA patients. Since Likert responses are easier to administer and interpret, they may be preferable to measure OA response.

Key words: Osteoarthritis, Visual analog scale, Likert rating scale, Endpoint correlation, Endpoint precision, Endpoint relationship.

Introduction

The evaluation of therapeutic efficacy in osteoarthritis (OA) patients is principally based on subjective rating scale responses to questions about pain, stiffness, function, and overall evaluation. The visual analog (VAS) and Likert scales are two major methods of measuring response to subjective questions about OA efficacy. Both methods of measurement were used in several efficacy questions in a 6-week, double blind, parallel group study of rofecoxib vs placebo OA[1]. The purpose of this article is to describe the relationship between the two types of scales and compare their measurement characteristics.

In his text “Quality of Life Assessments in Clinical Trails”, Spilker[2] has written that both methods have their proponents; however, there is a lack of evidence supporting a rational choice between them. Guyatt *et al.*[3] compared a VAS with a 7-point Likert scale in the assessment of quality of life in chronic pulmonary disease. They showed comparable levels of improvement in quality of life with both methods, greater variability with the VAS, and recommended the 7-point Likert scale over the VAS because of its ease of use and interpretation. Streiner and Norman[4] are also critical of the VAS. They reported that (1) subjects had more difficulty with a VAS than with numerical or adjectival scales; (2) a single item VAS is likely to demonstrate low reliability; and (3) other methods may yield more precise measurement, and possibly increased satisfaction among respondents.

On the surface, it may appear that the VAS may have better precision and be more sensitive to detect change than Likert scales simply because of finer gradations of levels of response. Preferred use of VAS has been endorsed on this basis[4]. However, a large body of work does not support this notion. Bellamy *et al.*[5,6] found similar relative efficiency (as a measure of precision) of the 100 mm VAS vs 5 point Likert scale responses using

Financial support for this work was from Merck Research Laboratories, Rahway, NJ 17065.

*Address correspondence and reprint requests to: James A. Bolognese, Merck Research Labs, P.O. Box 2000, Rahway, NJ 07065, USA. Tel: 732-594-4412; Fax: 732-594-6075; E-mail: james_bolognese@merck.com

Received 27 February 2003; revision accepted 21 March 2003.

within-treatment changes from baseline of the WOMAC Arthritis Index in patients with OA of the hip or knee. Jensen *et al.*[7] found similar magnitudes of correct response and predictive validity of the following six rating scales in chronic pain patients: 100 mm VAS, 0–100 point and 0–10 point numeric rating scales, and 4-, 5-, and 6-point Likert scales. Downie *et al.*[8] compared the measurement error of a vertical and horizontal 100 mm VAS to that of a 4 point Likert scale and a 0–10 point numeric rating scale. They found the highest measurement error with the 4 point Likert scale, and the lowest with the 0–10 point numeric rating scale, with the VAS scales midway between them. Bellamy *et al.*[9,10] found a slight numeric advantage of the VAS over the Likert scale in terms of effect size for pain measurement in OA (one article) and in rheumatoid arthritis (second article); however, both showed slight numeric advantage over the McGill pain questionnaire.

Regarding the use of Likert scales, Streiner and Norman[4] claimed that respondents are unable to discriminate much beyond 7 levels. They reported that descriptions of each level of response are preferred over descriptive anchoring at only the ends because the latter tend to increase variability by pulling responses towards the ends. Response scales using only positive integers are preferable to those using both negative and positive integers because respondents tend to choose only positive numbered categories[4]. Unless the distribution of scores is severely skewed, one can analyze data from Likert scales as if they were interval without introducing severe bias.

Hence, there is the suggestion that use of Likert scales may be as good as VAS, and almost certainly not substantially worse. All of the previous work was based on single respondent observations or on within-respondent changes from baseline. The present report extends this work to a placebo-controlled trial and centers the precision comparison on change from baseline difference between active treatment and placebo. It further assesses the potential usefulness of percent change and log transformation in the analysis of VAS responses, and assesses the relationship between the two scales by displaying specific magnitudes of VAS changes for categories of Likert change from baseline.

Materials and methods

Each of three OA efficacy questions were assessed using both a 0–100 mm VAS and a 5 point Likert scale in a 6-week, double-blind, parallel group study of the effect of rofecoxib vs placebo in 219 OA patients[1] treated with placebo ($N=72$), rofecoxib 25 mg once daily ($N=73$), or rofecoxib 125 mg once daily ($N=74$). The OA efficacy questions and their respective VAS and Likert response sets are described below; both response sets were used in this Phase II trial (the first OA study of rofecoxib) to assess their relative performance and support choices of endpoints for Phase III trials. Each question was administered at the prestudy visit (while patients were taking prior NSAID therapy for OA), the baseline visit (after OA flare, i.e., meeting pre-specified criteria for worsening of OA symptoms during washout from prior NSAID therapy), treatment Weeks 1, 2, 4, and 6 visits, and at the discontinuation visit if the patient discontinued the study prematurely.

PATIENT PAIN ASSESSMENT (VAS)

“The following question concerns the amount of pain due to arthritis in your study knee”.

VAS

“Please indicate the amount of pain recently experienced by marking an (X) through the line:”

100 mm VAS scale—Left hand marker “no pain”, right hand marker “extreme pain”.

|--|

no pain extreme pain

Likert

“Please indicate the amount of pain recently experienced by marking an (X) in one box below:”

□ No pain
□ Mild pain
□ Moderate pain
□ Severe pain
□ Extreme pain

For analysis, the responses were assigned numeric values 0 through 4, respectively.

PATIENT GLOBAL ASSESSMENT OF DISEASE STATUS

VAS

“Considering all the ways your arthritis affects you, mark (X) on the scale for how well you are doing.” 100 mm VAS scale—Left hand marker “Very well”, right hand marker “Very poor”.

Likert

“Considering all the ways your arthritis affects you, mark an (X) in one box below for how well you are doing.” Very well, Well, Fair, Poor, Very poor. For analysis, the responses were assigned numeric values 0 through 4, respectively.

INVESTIGATOR GLOBAL ASSESSMENT OF DISEASE STATUS

VAS

“Make a global assessment of the patient’s disease status by marking an (X) on the line below.” 100 mm VAS scale—Left hand marker “Very well”, right hand marker “Very poor.”

Likert

“Make a global assessment of the patient’s disease status by marking an (X) in one box below.” Very well, Well, Fair, Poor, Very poor. For analysis, the responses were assigned numeric values 0 through 4, respectively.

WOMAC VA 3.0 OA INDEX QUESTIONNAIRE WITH A 0–100 MM VAS RESPONSE SET

The WOMAC Pain Subscale asks “How much pain do you have?” for each of five situations; each was rated on a 0–100 mm VAS. The five situations are walking on a flat surface, going up or down stairs, at night while in bed, sitting or lying, and standing upright[5]. For purposes of this exercise, the WOMAC Pain Subscale (average of five pain questions) was also compared to the Likert scale Pain response because it is widely used and was the primary endpoint in the study which generated the data for this analysis.

In order to increase patients' understanding of VAS use, a standardized training video was shown to each patient prior to study start. This was not done with the Likert scale because it was felt that the Likert response set was self-explanatory. Thus, the comparisons in this article are between an enhanced VAS approach vs an unenhanced Likert approach to measuring OA efficacy.

STATISTICAL ANALYSES

The distribution (plots and summary statistics) of VAS scores at each timepoint was summarized within each baseline Likert scale category to examine the numeric relationship of Likert score to VAS score by time. Cumulative distribution plots were provided to display the percent of patients with values exceeding every observed VAS value; mean and median were computed to describe the location of the "center" of the distribution; and standard deviation (SD) and inter-quartile range (the difference between the 75th and 25th percentiles) were computed to quantify the spread of the data. These summaries were made for each Likert/VAS pair.

Change from baseline (at all on-treatment timepoints: Weeks 1, 2, 4, 6 and discontinuation) in each of the Likert scale variables mentioned in the Objectives was categorized (e.g., 4→4, 4→3, 4→2, 4→1, 4→0, 3→4, 3→3, etc., where 4, 3, 2, 1, and 0, represent the descriptors ("very well", "well", etc.) detailed in the previous section). The distribution of the associated VAS changes within each of these Likert change categories was summarized via box-plots, cumulative distribution plots, and summary statistics for each Likert/VAS pair of endpoints. The location and dispersion of the data were summarized by mean and SD, respectively. Although the cumulative distribution plots revealed the typical s-shaped curve approximating a normal distribution curve, median and inter-quartile range were also calculated to corroborate the relationships revealed by the means and SD's.

Pearson correlation coefficients between VAS and Likert endpoints for change from baseline and at baseline were computed to assess the strength of a linear relationship between the two scales. Coefficients from 0 to 0.2 indicate none or nearly no linear relationship, 0.2 to 0.4 indicate weak linear relationship, 0.4 to 0.6 indicate a moderate degree of linear relationship, 0.6 to 0.8 indicate strong linear relationship, and values from 0.8 to 1.0 indicate very strong linear relationship.

In order to assess whether the data from the pairs of Likert and VAS questions could be combined across endpoints to increase precision of the estimates of the relationship (i.e., the calibration of the VAS in terms of category of Likert scale response), analysis of variance (ANOVA) was performed to assess the dependent variable "VAS change from baseline" as a function of independent variables for endpoint, timepoint, and their interaction.

Percent change, and natural log of the on-treatment (A) vs baseline (B) ratio (ln(A/B)) was investigated to determine if more consistent relationships between VAS and Likert responses could be revealed in comparison to actual change from baseline. This was addressed by assessing random errors (i.e., residuals) from the above ANOVA for normality and variance homogeneity using the Shapiro–Wilk and Hartley's maximum variance ratio statistics, respectively.

During the assembly of the results, it was hypothesized that patients may become accustomed, or learn how to answer the VAS. To examine this, the variability of VAS scores was assessed in the small subset of patients whose category of Likert scale response did not change at the screening, and Week 1, 2, 4, and 6 visits. If the variability of VAS responses remained stable among these visits, then there would be no learning effect. However, if the variability would get smaller over time, then this could be a signal of some learning effect of the VAS. (Note that patients were not permitted to see their previous VAS responses.)

Results

DISTRIBUTION OF VAS SCORES BY LIKERT CATEGORY OF RESPONSE

Examination of the distributions of VAS scores for each Likert category of response at all the visits revealed generally s-shaped cumulative distributions indicating that they could be approximated by a normal distribution. Examples of these curves are shown for baseline Likert score of 3 in Fig. 1a through d. Hence, mean and SD were used to estimate location and variability (dispersion) parametrically. Median and inter-quartile range showed similar results as mean and SD, respectively.

At each individual timepoint, increments in VAS score between adjacent categories of Likert response ranged from 10 to 25 mm (Table I). In every case, mean VAS scores increased with increasing Likert score. Mean VAS values were generally similar in magnitude within each Likert value across baseline and weeks 1 to 6 with two exceptions. The first exception was larger mean VAS values at baseline within Likert values 1 and 2 only. For example, for Likert category 1, overall pain mean VAS score was 51 at baseline, and ranged from 20 to 26 across Weeks 1 to 6; for Likert category 2, mean at baseline was 63, and ranged from 53 to 58 at Weeks 1 to 6. The second exception was that screening mean VAS values were generally smaller than those at Weeks 1 to 6 within each Likert category 3 and 4; however, they were similar in categories 1 and 2, and larger in category 0, although the numbers of patients contributing are small at the extreme Likert categories 0 and 4. Summary statistics at both ends of the Likert scale are based on small sample sizes, and, therefore, may be unreliable. Refer to Table I for more detail.

SD's were generally similar across all timepoints within VAS categories, and generally smaller at the ends of the Likert scale within all timepoints. Similar relationships were seen for medians as were seen for means, and similar relationships were seen for inter-quartile ranges as were seen for SD's (data not shown).

DISTRIBUTION OF VAS SCORE CHANGES FROM BASELINE FOR EACH CATEGORY OF LIKERT SCALE CHANGE FROM BASELINE

Figure 2a through d show box-plots of the distribution of VAS changes for each category of Likert pre/post combination. The relationship between the VAS and Likert changes are summarized by median and inter-quartile range as shown in box-plots, and by mean and SD as listed in Table II. With regard to location, as shown in Fig. 2a through d and Table II, in every case, the magnitude of change

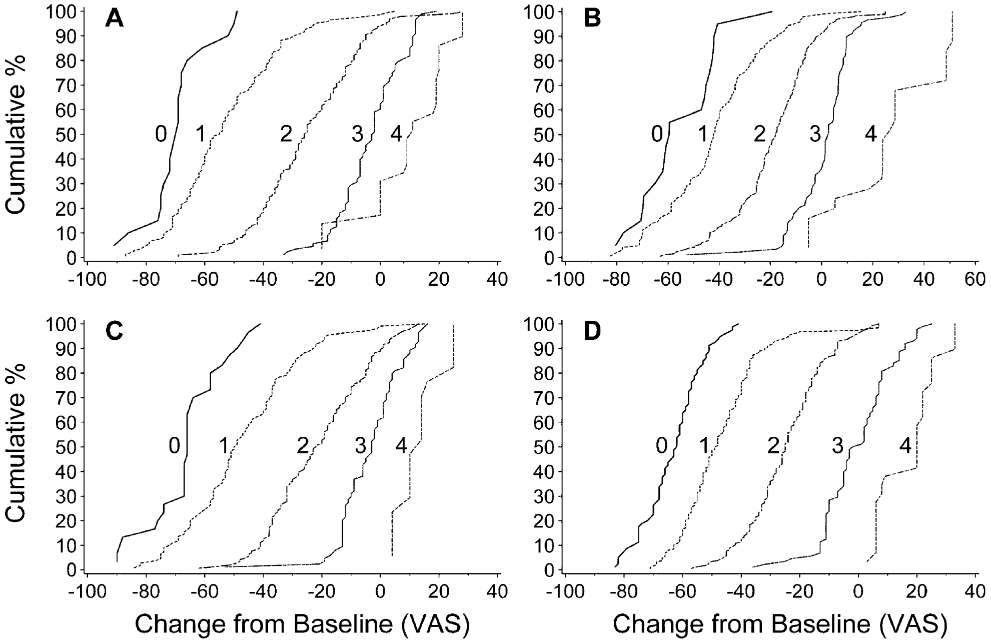

Fig. 1. Cumulative distributions of VAS changes from baseline for on-treatment Likert scale responses of 0, 1, 2, 3, and 4 from baseline Likert scale response 3 (Panel A, Overall Pain; Panel B, WOMAC Pain Subscale; Panel C, Patient Global Assessment of Disease Status; Panel D, Investigator Global Assessment of Disease Status).

Table I
Mean±SD (N) of VAS values (mm) by likert value at each visit

Likert value	Week −1	Week 0	Week 1	Week 2	Week 4	Week 6
Investigator Global Assessment of Disease Status						
0	8±4 (11)	n/a	10±6 (19)	8±4 (33)	9±5 (42)	8±5 (48)
1	24±11 (73)	40±14 (8)	25±11 (69)	24±11 (82)	23±10 (69)	24±11 (70)
2	47±12 (105)	58±11 (48)	49±12 (72)	45±11 (53)	50±14 (53)	49±11 (42)
3	66±11 (15)	73±8 (137)	73±9 (37)	75±10 (32)	75±9 (34)	74±9 (38)
4	64± (1)	84±7 (16)	91±7 (6)	90±10 (9)	92±6 (11)	92±6 (11)
Overall pain question						
0	34±32 (2)	n/a	4±2 (8)	7±9 (12)	7±7 (12)	5±4 (27)
1	27±15 (75)	51±6 (6)	26±18 (70)	21±15 (77)	20±16 (84)	21±14(68)
2	48±15 (103)	63±14 (66)	56±17 (81)	57±16 (85)	58±16 (67)	53±18 (75)
3	62±12 (23)	80±11 (120)	79±10 (35)	81±9 (27)	76±14 (36)	79±12 (30)
4	59±16 (2)	92±5 (18)	90±10 (9)	90±10 (8)	91±9 (10)	91±9 (9)
Patient Global Assessment of Disease Status						
0	11±12 (7)	n/a	8±13 (14)	9±15 (19)	5±5 (22)	5±5 (32)
1	27±16 (57)	41±16 (15)	28±16 (68)	26±15(75)	21±14 (73)	23±14 (73)
2	48±13(116)	58±15 (73)	52±15 (73)	50±16 (76)	54±15 (73)	51±16 (63)
3	63±14 (19)	73±12 (98)	73±12 (40)	73±17 (29)	78±12 (30)	77±14 (30)
4	77±15 (4)	89±6 (21)	94±4 (6)	95±2 (7)	95±2 (8)	91±11 (8)
WOMAC Pain Subscale						
0	15±8 (2)	n/a	4±2 (8)	7±6 (12)	6±4 (12)	7±6 (27)
1	27±13 (74)	17±7 (6)	24±15 (70)	20±13 (77)	19±14 (84)	20±13 (68)
2	46±16 (102)	51±18 (66)	48±16 (80)	48±16 (84)	50±16 (66)	46±18 (74)
3	57±17 (23)	67±16 (119)	71±12 (36)	67±17 (27)	67±18 (36)	71±16 (30)
4	67± 6 (2)	75±15 (18)	81±12 (8)	82±13 (7)	82±11 (9)	83±12 (8)

from baseline in VAS increased with increasing change from baseline for each category of Likert score change from baseline. With regard to spread in the data, middle categories of Likert scale change from baseline tended to spread more than those at the extremes. For example, the SD of VAS change from baseline was 10 for category of Likert change from 3 at baseline to 0 on treatment and to 4 on treatment in comparison to SD's of 14, 13, and 12 for category of Likert change from 3 at baseline to 1, 2, and 3 on treatment, respectively. Similar patterns were seen for other endpoints and other categories of Likert changes. However, in nearly all cases, there was a wide range of

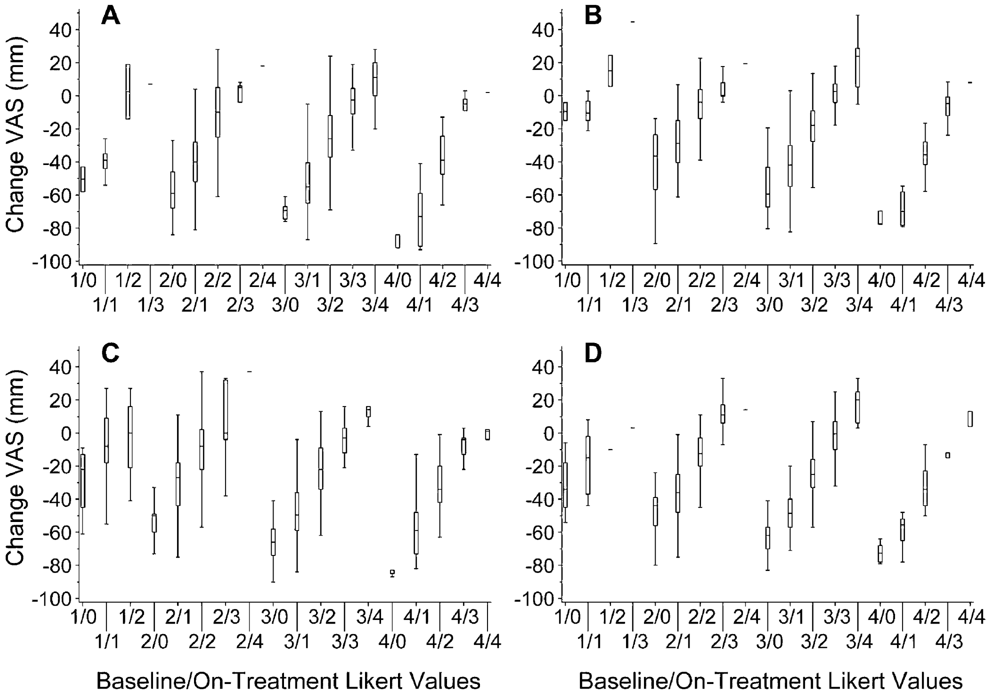

Fig. 2. Box-plots of distributions of VAS changes from baseline for all combinations of Likert scale baseline and on-treatment responses (Panel A, Overall Pain; Panel B, WOMAC Pain Subscale; Panel C, Patient Global Assessment of Disease Status; Panel D, Investigator Global Assessment of Disease Status). [Definitions of the Box-plots: the ends of the boxes represent the 25th and 75th percentiles; the horizontal line inside the box represents the median (50th percentile); the vertical lines ending at 'T' represent distance from the median of 1.5 times the difference between the 25th and 75th percentiles; all data except outliers should fall within the 'T's'; outliers would have been shown as '*' lying beyond the 'T's'.]

VAS changes associated with each level of Likert change for each Likert baseline value.

ANALYSIS AIMED AT COMBINING CATEGORIES OF LIKERT SCALE CHANGE FROM BASELINE

Table II shows the mean VAS changes from baseline for each Likert category of change from baseline for each endpoint. For each category of Likert change from baseline, ANOVA results indicated significant differences for mean VAS changes depending on baseline Likert score; hence, VAS changes cannot be combined across Likert baseline categories to assess relationship to corresponding Likert changes. For example, for Investigator Assessment of Disease Status, mean VAS changes for a Likert change of zero increased with increasing baseline Likert scores (means −18, −11, 0, and +9 for baseline Likert scores 1, 2, 3, and 4, respectively). A similar but less pronounced relationship was seen for Patient Assessment of Disease Status, WOMAC Pain Subscale, and Overall Pain Assessment. Thus, for the less severe baseline Likert categories (1 and 2), no change in Likert response was, on average, associated with some level of improvement detected by VAS.

For all four endpoints, mean VAS changes for a Likert improvement of −1 decreased (in absolute value, i.e., showed less improvement) with increasing baseline Likert categories. For example, mean changes in Overall Pain were −51, −40, −25, and −7 for baseline Likert scores 1, 2, 3, and 4, respectively. The same general relationship was seen for Likert improvements of 2 units (i.e., a change of −2). This probably reflects more noticeable improvements when Likert scores change 1 or 2 units from lower baselines. Mean improvements of 3 Likert units are only possible for baseline Likert scores of 3 or 4; for these categories of Likert change, mean VAS changes were largest (see Table II). Generally similar relationships were seen in similar examinations of changes from screening to baseline values. This is further evidence for not being able to combine categories of Likert changes across Likert baselines.

CORRELATIONS BETWEEN VAS AND LIKERT SCORES

Pearson correlation coefficients for each endpoint were measured for change from baseline, at baseline, and at screening. To assess correlations for change from baseline, the data were averaged within each patient across the 6 week treatment period (the primary endpoint of the study), and also assessed for change from baseline to the last observed timepoint.

Correlation coefficients for average change from baseline across the 6-week treatment period between VAS and Likert scales were 0.84, 0.76, and 0.73 for Investigator Global, Patient Global, and Overall Pain, respectively; the correlation between WOMAC Pain Subscale on the VAS and Overall Pain on the Likert Scale was 0.68. For change from screening to baseline, these values were smaller, 0.68, 0.46, 0.37, and 0.13, respectively. For change from baseline to last observed on-treatment value up to Week 6, these values were 0.85, 0.76, 0.74, and 0.71, respectively. Correlation coefficients at baseline were 0.72, 0.68, 0.66, and 0.38, respectively. Thus, except for the change from screening to baseline visit, there were strong correlations

Table II
Summary statistics of change in VAS (mm) for categories of change in Likert scale response

Likert values			Mean±SD (*N*)	Mean difference from treatment Likert value 0	Median [iqr]	Median difference from treatment Likert value 0
Baseline	Treatment	Change				
Overall pain question						
1	0	−1	−51±11 (2)	n/a	−51[15]	n/a
	1	0	−40±7 (19)	11	−39[9]	12
	2	+1	3±23 (2)	42	3[33]	42
	3	+2	7± (1)	5	7[0]	5
	4	+3				
2	0	−2	−56±15 (34)	n/a	−59[23]	n/a
	1	−1	−40±17 (125)	17	−40[25]	19
	2	0	−10±21 (87)	29	−10[30]	30
	3	+1	−1±16 (14)	9	5[15]	15
	4	+2	18±0 (3)	19	18[0]	13
3	0	−3	−69±10 (20)	n/a	−70[8]	n/a
	1	−2	−53±18 (144)	16	−55[25]	15
	2	−1	−25±17 (191)	28	−26[25]	29
	3	0	−3±11 (88)	21	−3[16]	24
	4	+1	9±15 (29)	13	11[20]	14
4	0	−4	−87±5 (3)	n/a	−84[8]	n/a
	1	−3	−73±18 (11)	13	−73[32]	11
	2	−2	−37±15 (28)	36	−39[25]	34
	3	−1	−7±11 (25)	30	−5[7]	34
	4	0	2±0 (4)	9	2[0]	7
WOMAC Pain Subscale						
1	0	−1	−10±8 (2)	n/a	−10[11]	n/a
	1	0	−9±7 (19)	0	−11[12]	-1
	2	+1	15±13 (2)	24	15[19]	26
	3	+2	45± (1)	30	45[0]	30
	4	+3	(0)			
2	0	−2	−39±20 (34)	n/a	−37[34]	n/a
	1	−1	−28±17 (125)	12	−29[27]	8
	2	0	−6±13 (87)	22	−4[18]	25
	3	+1	1±14 (14)	6	0[11]	4
	4	+2	19±0 (3)	19	19[0]	19
3	0	−3	−55±15 (20)	n/a	−60[25]	n/a
	1	−2	−42±19 (144)	13	−42[25]	17
	2	−1	−19±16 (187)	23	−18[19]	24
	3	0	1±12 (89)	20	3[12]	21
	4	+1	27±20 (25)	26	27[37]	24
4	0	−4	−75±5 (3)	n/a	−77[8]	n/a
	1	−3	−68±10 (11)	7	−70[20]	7
	2	−2	−36±17 (28)	32	−36[15]	34
	3	−1	4±25 (25)	40	−5[20]	31
	4	0	8±0 (4)	4	8[0]	13
Patient Global Assessment of Disease Status						
1	0	−1	−30±18 (14)	n/a	−22[33]	n/a
	1	0	−7±19 (35)	22	−8[27]	14
	2	+1	−2±20 (11)	5	0[37]	8
	3	+2	(0)			
	4	+3	(0)			
2	0	−2	−52±13 (38)	n/a	−50[12]	n/a
	1	−1	−30±18 (123)	21	−27[26]	23
	2	0	−9±20 (104)	21	−8[24]	19
	3	+1	6±21 (23)	15	0[36]	8
	4	+2	37±0 (3)	31	37[0]	37
3	0	−3	−66±13 (30)	n/a	−66[16]	n/a
	1	−2	−47±19 (110)	19	−50[23]	17
	2	−1	−21±16 (144)	26	−22[25]	28
	3	0	−3±11 (84)	18	−3[15]	19
	4	+1	13±8 (17)	17	14[14]	17
4	0	−4	−76±21 (5)	n/a	−85[25]	n/a
	1	−3	−57±19 (21)	19	−59[27]	26

Table II
Continued

Likert values			Mean±SD (*N*)	Mean difference from treatment Likert value 0	Median [iqr]	Median difference from treatment Likert value 0
Baseline	Treatment	Change				
	2	−2	−33±16 (26)	24	−34[23]	25
	3	−1	−10±13 (22)	23	−4[12]	30
	4	0	−4±10 (9)	6	1[6]	5
Investigator Global Assessment of Disease Status						
1	0	−1	−31±16 (17)	n/a	−34[30]	n/a
	1	0	−18±18 (13)	13	−15[38]	19
	2	+1	−10± (1)	8	−10[0]	5
	3	+2	3± (1)	13	3[0]	13
	4	+3				
2	0	−2	−48±14 (39)	n/a	−44[17]	n/a
	1	−1	−37±17 (77)	11	−36[23]	8
	2	0	−11±14 (42)	26	−13[17]	24
	3	+1	11±11 (30)	22	11[13]	24
	4	+2	14±0(3)	3	14[0]	3
3	0	−3	−63±10 (80)	n/a	−62[13]	n/a
	1	−2	−47±14 (188)	16	−49[17]	14
	2	−1	−25±13 (158)	22	−25[17]	24
	3	0	−0±12 (88)	24	−1[17]	25
	4	+1	18±10 (29)	18	20[19]	21
4	0	−4	−72±6 (6)	n/a	−73[11]	n/a
	1	−3	−59±9 (12)	14	−56[15]	17
	2	−2	−33±13 (19)	26	−34[21]	22
	3	−1	−12±6 (22)	21	−12[3]	22
	4	0	9±5 (5)	22	13[9]	25

iqr, inter-quartile range; i.e., 75th percentile minus 25th percentile.

between VAS and Likert scale responses in measuring the three endpoints and between the WOMAC Pain Subscale (VAS) and Overall Pain (Likert). Correlations for change from screening to baseline were modest to strong, except for the WOMAC and Overall Pain variables. These exceptions could be due to the homogeneity of patient response to withdrawal of NSAID; i.e., since patients were all worsening similarly, it was more difficult to demonstrate high correlation in such a narrow range of response. That is, a small variation occurring in a narrow range would appear as a small correlation.

ANALYSIS AIMED AT COMBINING ENDPOINTS (TO ASSESS VAS/LIKERT RELATIONSHIP WITH MORE PRECISION)

For each of the Likert and VAS response scales, the change from baseline data at Weeks −1 (Screening), 1, 2, 4, and 6 for the Overall Pain, Patient Global, and Investigator Global variables were combined for analysis to assess whether similar changes were observed across the various timepoints and across the variables. If they would be found similar, then the endpoints could be combined for more precise estimation of the statistics summarizing the VAS relationship to Likert. The mean changes from baseline were assessed via ANOVA with independent variables endpoint and Week.

The interaction between endpoint and Week was tested and did not approach significance (p>0.5), so differences among endpoints were consistent across Weeks, and vice versa. This justifies the combination of the data across study weeks to generate the data summaries presented thus far. Mean change from baseline differed significantly across endpoints. Hence, combining the endpoint distributions to assess the VAS/Likert relationship is not clearly straightforward since their distributions were shifted and since these shifts among endpoints were different on the Likert and visual analog scales.

ASSESSMENT OF PERCENT CHANGE AND NATURAL LOG OF ON-TREATMENT TO BASELINE RATIO FOR DESCRIBING THE VAS/LIKERT RELATIONSHIP

The magnitudes of percent changes in VAS from baseline to Study Weeks 1, 2, 4, and 6 were compared with corresponding discrete Likert scale changes. (Note that it does not seem to make good sense to assess Likert Scale percent changes because of their discrete nature.)

The distributions of the actual VAS changes were closer to normality than those of percent change and of natural log (of on-treatment/baseline ratio) for each endpoint. Also, the variances of the actual VAS changes were more homogeneous than those of percent change and of natural log (of on-treatment/baseline ratio) for each endpoint.

In general, where there was sufficient sample size to yield reliable estimates, the differences between mean percent changes in VAS response for adjacent Likert scale categories were mostly between 20 and 40 percentage points (ranged from 12 to 70). Differences in median percent change responses were generally similar. (Detailed summary tables are not included to conserve space.)

Like the results for actual change from baseline, ANOVA of VAS percent change as a function of baseline showed significant differences among the baselines at most values of Likert change for each endpoint. Thus, the mean VAS percent change cannot be computed in a combined fashion across baseline values for the same value of Likert change.

Table III
Effect sizes† for VAS and Likert scale responses

Change from baseline	Scales	Overall pain	Patient Global	Inv. Global	WOMAC Pain
Average across treatment period	VAS	1.08	1.15	1.33	1.04
	Likert	1.05	1.10	1.28	Not measured
Last observed value over 6 weeks	VAS	0.89	1.00	1.09	0.89
	Likert	0.86	0.88	1.04	Not measured

†Mean difference between combined rofecoxib groups vs placebo divided by the pooled SD.

Table IV
Average and median SD's of all, last 3, and last 2 VAS observations for those patients with all Likert responses the same at all visits except baseline (Weeks −1, 1, 2, 4, 6)

Endpoint	Likert category	Number of patients†	Average SD across timepoints			Median SD across timepoints		
			All	Last3	Last2	All	Last3	Last2
Investigator Global Assessment of Disease Status	1	4	4.9	3.7	2.7	4.4	3.4	1.8
	2	7	8.9	7.2	6.6	11.2	5.5	4.9
Overall pain question	1	11	11.3	9.6	8.9	8.5	7.1	7.1
	2	11	12.7	10.3	9.4	11.4	8.7	4.9
Patient Global Assessment of Disease Status	1	9	8.2	7.2	6.8	7.2	6.0	5.7
	2	9	10.6	9.0	8.2	10.8	8.1	8.5
Womac Pain Subscale	1	11	7.8	4.7	3.6	6.3	3.5	1.8
	2	11	11.5	7.4	4.4	11.4	6.5	4.2

†Number of patients with the same category of Likert scale response at all visits except the baseline (Weeks −1, 1, 2, 4, and 6).

EFFECT SIZES OF VAS AND LIKERT RESPONSE SCALES

The relative utility of the VAS and Likert response scales in clinical trials was assessed by their effect sizes (the magnitude of difference between active treatment and placebo divided by the pooled SD; this expresses the magnitude of difference in units of SD). Table III shows the effect sizes for the difference between the combined rofecoxib doses (because they demonstrated similar levels of efficacy[1]) vs placebo in the trial. The effect sizes were generally similar in magnitude for the VAS and Likert scale, but only slightly larger for VAS. Effect sizes within each variable were all larger for the overall average (across the 6-week treatment period) change from baseline in comparison to change from baseline to the last observed value.

POTENTIAL FOR LEARNING EFFECTS ON THE VAS

The average and median SD's of the last 3 and last 2 VAS observations were smaller than those of all observations for all endpoints and every category of Likert scale response for which the sample size exceeded three patients (Table IV), consistent with a learning effect for VAS responses over time.

Discussion

In this OA study, Likert scale responses are highly correlated with VAS responses, and both generate similar precision when comparing active treatment to placebo. In addition, the results suggest that when analyzing clinical trial data for treatment effects in OA, VAS response can be measured using simple difference from baseline. There was no gain in precision, in closeness to normality, or in closeness to homogeneity of variance if percent change or log of on-treatment to baseline ratio were used instead of difference from baseline. Finally, the results suggest additional ways in which efficacy data from OA trials can be assessed to compare the relative efficacy across treatments and the relative quality of the efficacy information across trials.

A change between adjacent categories of Likert scale response (i.e., of 1 unit) yields different VAS mean changes depending on the category of Likert baseline score—decreasing VAS improvements for increased severity of Likert baseline category. So it's uncertain what a specific change in VAS means for an individual patient. The closer to the middle of the scale, the more stable the relationship. Since the Likert response categories are labeled with words, they have face validity and the changes are well defined. The question of whether a one point Likert change from different baselines demonstrates the same magnitude of improvement is unknown; but this pitfall of the Likert scale exists for the VAS, as well.

For some Likert scores and some changes in Likert scores, the associated VAS values can vary across a very wide range, indicating large variability between patients' VAS responses for individual discrete Likert scale responses. Since the Likert scale responses are anchored by words, this variability may be detrimental to the precision of the VAS to discriminate between treatments. The potential gain in precision due to continuous measurement may be offset by this variability since effect sizes are generally similar between VAS and Likert scale endpoints. The VAS variability does decrease at the extremes of the Likert response scale, as is expected due to ceiling/floor effects.

Distributions of actual VAS values were closer to normality and homogeneity of variance than percent change and log-transformed VAS values. Hence there appeared to be no advantage of using percent change or log transformation over the actual change in VAS.

The Likert scale is easier to use than the VAS for two major reasons. First, it is easier for a patient to understand the check-boxes each associated with a word or phrase rather than marking an X on a continuous line anchored with words only at both ends. Second, the Likert is easier to score since each check-box results in a unique, easily read response item in comparison to the measurement necessary to use the VAS. This pitfall of the VAS can, however, be overcome by using a 0–100 numeric rating scale; others have shown the usefulness of such a numeric scale[7].

Another down-side to the VAS is that photocopying VAS response scales can change the length of the scale a few millimeters. This can lead to increased variability depending on how many times the copies are copied to make the blank case report forms. For these reasons, and since the Likert scale responses correlate highly with, and have similar precision as their corresponding VAS responses, Likert scales may be more useful in clinical trials than VAS's. Several others have demonstrated similar effect sizes for VAS and Likert scales[5–10].

The percent of patients with large changes may be useful in discriminating between active treatments because relationships between VAS and Likert scales are clearer (i.e., tighter). Changes between adjacent categories of Likert scale response (1 unit) have large variability—changes of 2 units yield tighter variability. For example, from Table II, for the overall pain question, the ratio of VAS SD to mean change from baseline (inverse of within-treatment group effect size) was 1.6, 0.41, 0.25, and 0.06 for categories of Likert changes 4 to 3, 4 to 2, 4 to 1, and 4 to 0, respectively; was 0.68, 0.34, and 0.14 for Likert changes 3 to 2, 3 to 1, and 3 to 0, respectively; and was 0.43 and 0.27 for Likert changes 2 to 1 and 2 to 0, respectively. Most of these reductions were due to increases in mean VAS improvement, but some of it was due to tighter variability (smaller SD's) for the larger improvements. Similar relationships were seen for the other endpoints. However, the largest percentage point differences between the active treatment and placebo groups were produced by using moderately large VAS cutoffs of improvement. Depending on the endpoint, cutoffs of 11–18 mm of VAS improvement led to the largest percentage point differences between the active and placebo treatments for most endpoints examined. However, in some cases improvements as large as 40 mm yielded maximum percentage point differences (data not shown).

The effect sizes were generally similar in magnitude for the VAS and Likert scale; thus each should provide similar precision for assessing OA efficacy. Effect sizes within each variable were all larger for the overall average (across the 6-week treatment period) change from baseline in comparison to change from baseline to the last observed value. This is not surprising since use of the average tends to decrease variability (because it incorporates information from multiple assessments), and the level of treatment response was generally similar across time[1]. Thus, the average response tends to integrate the treatment differences observed at each week.

Smaller SD's were seen at later time points than earlier time points, consistent with some learning effect on VAS scores. Further study is required to assess if this is a detriment to using VAS scores.

Examination of the relationship between Likert and VAS responses may be useful to assess the internal consistency of efficacy results within a trial and to compare quality of efficacy results across trials. Within a trial, the degree of association between VAS and Likert responses could be used as a measure of internal consistency of efficacy results. Across Trials, Likert global response may be useful as an anchor to measure goodness of trial based on the degree of spread found in the VAS endpoints. That is, those trials with VAS results which most closely "match" (more consistent mean and less SD) the Likert global response anchor may be the best trials. This is also a function of inclusion/exclusion criteria which could produce a more or less homogeneous group of patients studied.

In summary, VAS and Likert scale responses are highly correlated and yield similar precision for discriminating active from placebo treatment in OA patients. Since Likert scale responses are easier to administer and interpret, it may be preferable to use them to measure OA response. Although not assessed in this study, a 0–10 point discrete scale may be the most useful compromise, incorporating all positive attributes of both the visual analogue and Likert scale responses; however, this requires further study.

Acknowledgements

The authors thank Dr Ned Braunstein for his review of the manuscript, helpful suggestions and encouragement.

References

1. Ehrich EW, Schnitzer TJ, McIlwain H, Levy R, Wolfe F, Weisman M, *et al.* Effect of specific COX-2 inhibition in osteoarthritis of the knee: a 6-week double-blind, placebo-controlled, pilot study of rofecoxib. J Rheum 1999;26:2438–47.
2. Spilker B. Quality of Life Assessments in Clinical Trails. New York: Raven Press 1990 pp. 53–54.
3. Guyatt GH, Townsend M, Berman LB, Keller JL. A comparison of Likert and Visual Analogue Scales for measuring change in function. J Chronic Dis 1987; 40:1129–33.
4. Streiner DL, Norman GR. Health Measurement Scales: a Practical Guide to their Development and Use. 2nd edn. New York: Oxford University Press 1995 pp. 33–39.
5. Bellamy N, Buchanan WW, Goldsmith CH, Campbell J, Stitt LW. Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes to anti-rheumatic drug therapy in patients with osteoarthritis of the hip or knee. J Rhematol 1988;15:1833–40.
6. Bellamy N, Buchanan WW, Goldsmith CH, Campbell J, Stitt L. Validation study of WOMAC: a health status instrument for measuring clinically-important patient-relevant outcomes following total hip or knee arthroplasty in osteoarthritis. J Orthop Rheum 1988; 1:95–108.
7. Jensen MP, Karoly P, Braver S. The measurement of clinical Ping intensity: a comparison of six methods. Pain 1986;27:117–26.
8. Downie WW, Leatham PA, Rhind VM, Wright V, Branco JA, Anderson JA. Studies with pain rating scales. Ann Rheum Dis 1978;37:378–81.
9. Bellamy N, Campbell J, Syrotuik J. Comparative study of self-rating pain scales in osteoarthritis patients. Curr Med Res Opin 1999;15:113–9.
10. Bellamy N, Campbell J, Syrotuik J. Comparative study of self-rating pain scales in rheumatoid arthritis patients. Curr Med Res Opin 1999;15:121–7.