Volume 22 • Number 8 • October 1999
0161-4754/99/$8.00 + 0 **76/1/101187** © 1999 JMPT

ORIGINAL ARTICLES

The Bournemouth Questionnaire: A Short-form Comprehensive Outcome Measure. I. Psychometric Properties in Back Pain Patients

Jennifer E. Bolton, PhD,[a] and Alan C. Breen, DC, PhD[a]

ABSTRACT

Objective: Develop and test a short-form comprehensive outcome measure for back pain.

Design: Prospective longitudinal study of 3 consecutive cohorts of back pain patients.

Setting: Anglo-European College of Chiropractic outpatient clinic and several field chiropractic practices.

Method: Domains judged important in the back pain model and responsive to clinical change were identified from the literature. Items were scored on an 11-point numerical rating scale. The instrument was psychometrically tested by use of those tests relevant to an evaluative measure.

Results: Seven dimensions of the back pain model were included in the questionnaire. Having established face validity, the instrument was shown to demonstrate high internal consistency (Cronbach's alpha = 0.9) and good test-retest reliability (ICC = 0.95). All items were retained on the basis that they contributed to the overall score (item-corrected total score correlations) and to the instrument's responsiveness to clinical change (item change-corrected total change score correlations). The instrument demonstrated acceptable construct and longitudinal construct validity with established external measures. The effect size of the instrument was high (1.29) and comparable with established measures.

Conclusion: A reliable, valid, and responsive instrument has been developed for use in back pain patients. It is practical for use in investigations of both the efficacy and effectiveness of back pain treatments. (J Manipulative Physiol Ther 1999;22:503-10)

Key Indexing Terms: Back Pain; Outcome Measure; Psychometric Testing

INTRODUCTION

As a result of the present climate of systematic reviews, clinical guidelines, and evidence-based practice, the methods by which the efficacy of treatment and health care are evaluated are the subject of intense attention and focus. Evidence-based practice is now considered as the way forward in ensuring that cost-effective treatment is delivered to most patients.[1,2] Because much of this evidence is based on findings from research trials, and in particular random controlled trials, there are now calls to determine treatment effectiveness in routine practice settings as well because outcomes here may not necessarily be the same as those in the research setting.[3-5] Whether in research or practice settings, however, the ways in which patient outcomes are measured is a central issue in the decision-making process of future treatment and health care regimens.

Because of the implications of back pain on patients and society at large, this complaint has been the subject of systematic reviews[6,7] and clinical guidelines.[8] As part of the process, it is essential that back pain outcomes are measured in an appropriate manner, reflecting the nature of the complaint and being both meaningful and relevant, particularly to the patient.[9] Selecting outcome measures for use in research trials in conditions such as back pain, however, has always been problematic because no incontrovertible or direct measures of back pain exist. The reason for this is simple. Pain, the primary symptom of back pain, is a multidimensional, individual experience or behavior with a number of sensory, affective, cognitive/behavioral, and social aspects.[10] The complex and subjective nature of pain is presented in Loeser's model of pain, in which nociception at the core is surrounded by successive layers of pain "experience and suffering."[11] Although acute pain may be considered to involve a mainly nociceptive response to tissue damage or injury and chronic pain to involve psychologic and behavioral mechanisms in addition to the physical,[11] pain of whatever origin or duration is a complex phenomenon necessitating a multidimensional approach.

Despite the complexity of pain generally, and back pain specifically, some agreement over the way in which back pain outcomes should be measured does exist. Certainly pain itself, as the predominant symptom, is invariably included in outcomes assessment in back pain patients. For this there are a number of pain scales, measuring both its quantitative[12] and qualitative[13] expressions. Functional status, in terms of day-to-day living and work activities, is arguably the most relevant outcome to the patient. Accordingly, a growing number of functional status instruments are available for use in the back pain patient.[14-19] In addition to these, there is a whole battery of instruments, both generic and back pain-specific, measuring psychologic influences, cognitive-behavioral aspects,

[a]Anglo-European College of Chiropractic, Bournemouth, England.
Submit reprint requests to: Dr J. E. Bolton, Anglo-European College of Chiropractic, 13-15 Parkwood Rd, Bournemouth BH5 2DF, England; *jbolton@aecc-chiropractic.ac.uk*.

Paper submitted December 21, 1998; in revised form January 19, 1999.

social roles, well-being and quality of life, overall improvement, and satisfaction with care, all of which have been used to measure outcomes in back pain patients.[20,21]

Those interested in evaluating treatment outcomes in back pain research trials, as well as those who wish to document patient outcomes in routine clinical practice, are therefore faced with a stark choice when selecting appropriate measures to use in their patients. On the one hand, they can choose a whole battery of measures, secure in the knowledge of covering all aspects of this conceptually complex complaint that are generally impractical. On the other hand, they can select a few simple measures that, although practical and easy to use, may not reflect the true complexity of the complaint and so risk the chance of results that are neither relevant nor meaningful. Moreover, with such a battery of measures to choose from, little standardization of outcome measures in trials evaluating back pain health care and, as a result, little opportunity for direct comparisons between treatment effects remains.

In light of these tensions, an urgent need exists for multidimensional and comprehensive outcome measures relevant to the back pain experience. Although a number of generic multidimensional health status instruments exist, for example the Sickness Impact Profile (SIP)[22] and the Short-Form 36 (SF-36),[23] few multidimensional measures are condition-specific and measure the wide variety of aspects of back pain behavior in a single instrument. On the basis of a review of the literature, a number of domains have been identified as important in the back pain model, for which a growing number of outcome measures have been developed and are in use at present.[9] However, these existing measures, even though multi-item, do not comprehensively encompass all the important domains of the back pain experience. Moreover, they are long and cumbersome, often requiring both time and expertise to administer and interpret. As a result, we have identified a need for a short-form, multidimensional back pain measure suitable for use in documenting outcomes both in the routine clinical setting and in clinical trials.

The aim of this study was therefore to develop a clinically useful, multiaspect outcome measure that was (1) comprehensively based on the conceptual model of back pain; (2) brief, in that each aspect was measured on a single-item global scale; (3) suitable for use in ambulatory back pain patients typical of those attending chiropractic outpatient clinics; (4) quick and easy to use in routine clinical practice and research-based settings; (5) acceptable to patients, clinicians, and researchers; and (6) reliable, valid, and responsive to clinically significant change.

METHODS

Questionnaire Development and Testing

Questionnaire development and testing were based on the methodologic frameworks outlined by Kirshner and Guyatt[24] and Streiner and Norman[25] for developing and assessing health indices.

Item selection. To generate a list of items to be included in the questionnaire, the literature was reviewed to ascertain those traits considered conceptually important in the back pain model. As evidence of this, it was decided to concentrate on those aspects of the back pain experience that were most commonly measured and shown to be responsive to clinically significant change. Seven aspects of the back pain experience were selected for inclusion in the questionnaire. These were the sensory component of pain intensity; functional status in terms of day-to-day activity and social activity; the affective domains of anxiety and depression; and the cognitive/behavioral expressions of fear-avoidance beliefs about work activity and pain locus control.

Item wording and scaling. All the attributes included in the questionnaire were represented by a global scale, hence reducing the long-form, multi-item questionnaires on which they were based to a single item. Because of this, it was essential to ensure that the wording of each item was clear and unambiguous. Pilot studies in patients, practitioners, and researchers resulted in changes to the way in which the questions were asked so that they were clear, particularly to patients. An 11-point numerical rating scale (NRS) was used as the scaling response for each of the items in the questionnaire.

A distinction was made between the pretreatment and posttreatment questionnaires in that the pretreatment measure included demographic details, whereas the posttreatment measure included a 6-point global satisfaction with treatment scale and a 7-point global improvement scale.

Reliability, validity, and responsiveness. For an outcome measure to be clinically useful in both research and practice settings, it must be reliable, valid, and responsive to clinically significant change. Outcome measures assess longitudinal change within patients over time, and as such require different psychometric properties from those of either discriminative or predictive instruments. For an evaluative measure to function appropriately, it must tap areas relevant and responsive to change in health status (item selection), show stable intra-subject variation with insignificant variation between stable replicate measures (test-retest reliability), display a strong relationship between change scores and change scores in external measures over time (external longitudinal construct validity), and have the power to detect clinically important differences over time (responsiveness).[24]

Data Collection

In addition to preliminary testing of the instrument for face validity, the study was conducted in 3 phases.

Phase 1 (pilot study). The questionnaire was piloted in patients who were seen at a single field chiropractic practice. New patients or patients being seen with a new complaint ("new-old" patients) of back pain (with or without leg pain) were administered the questionnaire at the initial visit, on the second visit before treatment (test-retest), and by mail 4 to 6 weeks later.

Phase 2 (homogeneity and reliability). After the results of the first phase were obtained, a modified version of the questionnaire was administered to new and "new-old" back pain (with or without leg pain) patients coming to the college teaching clinic at their initial visit before seeing the intern.

On leaving the clinic the same day, patients completed a second questionnaire (test-retest). It should be noted that the initial visit is a screening interview and no treatment takes place. Patients were asked to complete a posttreatment questionnaire 4 to 6 weeks later.

Phase 3 (validity and responsiveness). The same questionnaire as that used in phase 2 was administered to new and "new-old" back pain patients (with and without leg pain) coming to a number of field chiropractic practices together with a modified Chronic Pain Grade (CPG) questionnaire,[26] the Revised Oswestry Disability Questionnaire (RODQ),[16] the DRAM,[27] the Pain Locus of Control (PLC) scale,[28] and the FABQ.[29] Patients completed the same battery of questionnaires approximately 4 weeks after treatment.

Data Analysis

Homogeneity and reliability. It was important that the scale was homogenous in that all of the items tapped different aspects of the same attribute, in this case back pain behavior, so that all the items in the questionnaire could be summed to form a total score. Items should be moderately correlated with each other and each should correlate with the total scale score. Homogeneity of the scale was tested using Cronbach's alpha coefficient and item-corrected total correlations.[25] Homogeneity is considered acceptable when Cronbach's alpha exceeds 0.7 but is not higher than 0.9, and item score correlations with corrected total scores are not less than 0.2.[25]

Test-retest reliability was investigated by use of the Intra-Class Correlation (ICC) coefficient. Because of an unacceptably long time between administrations of the test and retest questionnaires in phase 1 of the study (range, 1 to 37 days), patients could not be assumed to be stable. Therefore data from phase 2 of the study in which 2 administrations of the questionnaire were made on the same day were used instead. Only those patients who reported no change in their back pain on the global improvement scale between the 2 administrations were included in the analysis (78% of the sample, $n = 61$). To reduce the probability of memorizing their initial responses, items on the second questionnaire were given in a different order from that on the first questionnaire. ICC values (ranging from 0 to 1 on a continuum of increasing strength of agreement) were calculated from a repeated measures 2-way analysis of variance table.[25]

Validity. Face validity, in terms of acceptability to patients and researchers, was tested in a preliminary phase and in phase 1 of the study. Content validity, in terms of the completeness with which a questionnaire covers the important aspects of the attribute it is supposedly measuring,[24] was addressed at the outset of this study in selecting items to include in the questionnaire that were considered salient to the conceptual nature of back pain behavior and subject to clinically important change after treatment.

External construct validity is concerned with the extent to which a measure relates to other measures in a manner consistent with the theoretical constructs under test. In this respect, the correlations of items in the questionnaire and their counterpart external instruments were measured. Kirshner and Guyatt[24] have argued that "…for the most convincing, or definitive, demonstration of the validity of an evaluative instrument, its relation to other measures must be examined prospectively in a setting in which change over time is measured." Accordingly, longitudinal external construct validity was measured by correlating the within-subject longitudinal changes in questionnaire scores with those that use established measures. In all cases, construct validity was determined by use of Pearson's correlation coefficient (r)[25] in data obtained from phase 3 of the study.

Responsiveness. For a measure to be useful as an outcome measure it is imperative that it has the ability to detect clinically important changes over time.[30] Clinically significant changes remain difficult to define and do not necessarily equate with statistically significant changes. For the purposes of this study, clinically significant changes were defined as those in which patients themselves reported improvement in their complaint on a global improvement scale. Therefore only those patients who reported improvement were included in the responsiveness analyses on pretreatment and posttreatment scores taken from phase 3 of the study. This approach is advocated by Kirshner and Guyatt,[24] who describe the evaluation of instrument responsiveness by the examination of change scores in those who, by other criteria, improve or deteriorate. In this particular case, perceived improvement was the external criterion by which responsiveness of the questionnaire was assessed. With this approach, all patients, except 1, registered improvement on the global scale and were therefore included in the analysis ($n = 55$).

Internal responsiveness of individual items of the questionnaire was investigated by determining the strength of the correlations between the change score for each item and the corrected total change score (also known as internal longitudinal construct validity). This correlation should be 0.3 or greater to ensure that each item contributes significantly to the overall responsiveness of the instrument.[31]

External responsiveness was investigated by comparing the ability of the questionnaire to detect clinically significant change with that of external measures. Although several methods exist with which to calculate responsiveness,[30,32] the effect size as calculated by Kaziz et al[33] was used in this study. On the basis of the effect size, the number of subjects required to demonstrate this effect as statistically significant at the 5% level with 90% power was calculated, both for the questionnaire and the external measures used in phase 3 of the study.[34]

RESULTS

Subjects

Ninety patients were recruited to phase 1 of the study. The mean age of these patients was 52.1 (SD = 15.5) years with a higher proportion of female patients (62.2%). In phase 2 of the study, complete data were obtained from 82 patients in which the mean time period between pretreatment and posttreatment questionnaires was 44.9 (SD = 4.2) days. The mean age of this cohort was 49.5 (SD = 16.9) years; 53.7% of the sample were women. These data were used to calculate the internal consis-

Table 1. *Internal consistency of the Bournemouth Questionnaire*

	Item-corrected total correlations							Cronbach's alpha
	Pearson's *r*							
Item	1	2	3	4	5	6	7	Total score
Pretreatment	0.75	0.74	0.68	0.59	0.65	0.47	0.72	0.87
Retest	0.69	0.66	0.54	0.70	0.62	0.74	0.72	0.89
Posttreatment	0.76	0.70	0.72	0.60	0.63	0.72	0.76	0.91

Number of observations = 82.

Table 2. *External construct validity of items of the Bournemouth Questionnaire*

		Pearson's *r*	
Item	Correlation with	Pretreatment	Posttreatment
1	CPQ (average pain intensity scale)	0.78	0.89
2	MSPQ	0.41	0.61
3	CPQ (daily activities disability scale)	0.82	0.83
	RODQ	0.85	0.81
4	CPQ (ability to work scale)	0.55	0.58
	FABQ	0.56	0.66
5	Zung	0.57	0.70
6	PLC	0.40	0.49
7	CPQ (social activities disability scale)	0.54	0.82
	RODQ	0.58	0.66

Number of observations = 55.
CPQ, Chronic Pain Questionnaire; *RODQ,* Revised Oswestry Disability Questionnaire; *MSPQ,* Modified Somatic Pain Questionnaire; *FABQ,* Fear Avoidance Beliefs Questionnaire; *PLC,* Pain Locus of Control questionnaire.

tency of the questionnaire. Test-retest data were available in 61 of these patients. In phase 3 of the study, complete data were obtained from 55 patients, and the mean time period between completion of pretreatment and posttreatment questionnaires was 32.2 (SD = 14.7) days. The mean age of this cohort was 45.7 (SD = 12.5) years, and gender distribution was approximately equal (50.9% women). These data were used to test the validity and responsiveness of the questionnaire.

Pilot Study (Face Validity)

After revisions to the wording of questions in the preliminary phase of the study, the questionnaire piloted in phase 1 was found to be acceptable and relevant to patients; very few patients either marked the questionnaire to indicate that they had any difficulty in answering the questions or left questions blank. Even so, some minor revisions to the wording were made after phase 1 of the study. In this phase, the response scales to 3 of the 7 global dimensions in the questionnaire were "reversed" to deter patients from simply circling the same number on each scale without reading each question in full. This probably led, however, to some of the respondents misinterpreting the direction of the scale and, as a result, circling the wrong number. As a result, the questionnaire was modified in subsequent phases of the study and all items scored in the same direction.

The 7 global dimensions from the modified version of the questionnaire tested in phases 2 and 3 are given in Appendix 1. The questionnaire is referred to hereafter as the Bournemouth Questionnaire (BQ).

Homogeneity of Items

The BQ demonstrated good internal consistency, indicating that all the item scores can be summed to give a total score (Table 1). For each of the 3 administrations (ie, pretest, retest, and posttreatment) of the BQ in phase 2 of the study, Cronbach's alpha was approximately 0.9, supporting the hypothesis that the instrument taps different aspects of the same attribute. All the item-corrected total correlations were well above the cut-off of 0.2 advocated by Streiner and Norman,[25] indicating that all of the items contributed to the overall score and should therefore be retained.

Test-retest Reliability

The ICC of the total score of the BQ on the basis of 2 administrations in stable patients in phase 2 of the study was 0.95 (n = 61), indicating strong agreement between total scores in these patients. The limits of agreement (equal to the absolute mean difference between test-retest observations ±2 standard error of the differences) were 2.6 and 4.5, demonstrating that change scores greater than 4.5 are indicative of real change beyond the variability in change scores in stable subjects who used this scale.

Validity

Table 2 shows data obtained from phase 3 of the study testing the external construct validity between individual items of the BQ and external measures evaluating similar aspects of the back pain experience. BQ items were tested either against whole instruments or against individual items of other measures as appropriate (Table 2). Construct validity was tested for both the pretreatment and the posttreatment questionnaires. As shown, items of the BQ showed overall good construct validity, with comparative external measures supporting the construct validity of the scale. The correlation for the locus of control item was low for both the pretreatment and posttreatment questionnaires (0.40 and 0.49, respectively) as was the correlation for the anxiety item (item 2) on the pretreatment questionnaire. Nevertheless, all correlations were statistically significant (P <.05).

Tables 3 and 4 show the external longitudinal construct validity of the BQ against the established measures used in this study, again by use of data obtained in phase 3. Table 3 shows the correlations between change scores in individual items of the BQ and change scores in counterpart external measures as shown in Table 2. As can be seen, correlations between change scores in the pain intensity and physical activity scales of the BQ and their counterpart external measures were stronger than those between the items associated with cognitive and affective aspects of the back pain experience, which were generally moderate to weak. Nevertheless, all correlations were statistically significant apart from the correlation in change scores between item 4 of the BQ and its counterpart FABQ (Table 3).

Table 3. *External longitudinal construct validity of items of the Bournemouth Questionnaire*

Item	Correlation with	Pearson's *r*
1	CPQ (average pain intensity scale)	0.69
2	MSPQ	0.34
3	CPQ (daily activities disability scale)	0.78
	RODQ	0.79
4	CPQ (ability to work scale)	0.34
	FABQ	0.24
5	Zung	0.29
6	PLC	0.38
7	CPQ (social activities disability scale)	0.37
	RODQ	0.57

Number of observations = 55.

CPQ, Chronic Pain Questionnaire; *RODQ,* Revised Oswestry Disability Questionnaire; *MSPQ,* Modified Somatic Pain Questionnaire; *FABQ,* Fear Avoidance Beliefs Questionnaire; *PLC,* Pain Locus of Control questionnaire.

Table 4. *External longitudinal construct validity of the Bournemouth Questionnaire*

Correlation of BQ with	CPQ	RODQ	PLC	MSPQ	Zung	FABQ
Pearson's *r*	0.77	0.78	0.40	0.36	0.40	0.32

Number of observations = 55.

BQ, Bournemouth Questionnaire; *CPQ,* Chronic Pain Questionnaire; *RODQ,* Revised Oswestry Disability Questionnaire; *MSPQ,* Modified Somatic Pain Questionnaire; *FABQ,* Fear Avoidance Beliefs Questionnaire; *PLC,* Pain Locus of Control questionnaire.

Table 5. *Responsiveness (internal longitudinal construct validity) of items of the Bournemouth Questionnaire*

	Change score item-corrected total change score correlations						
	Pearson's *r*						
Item	1	2	3	4	5	6	7
	0.64	0.71	0.65	0.64	0.56	0.66	0.61

Number of observations = 55.

Table 6. *Effect sizes of the Bournemouth Questionnaire and other outcome measures*

	Effect size	*N*
BQ	1.29	9
CPQ	1.44	7
RODQ	1.07	11
PLC	0.60	32
MSPQ	0.59	32
Zung	0.57	37
FABQ	0.004	—

Number of observations = 55.

N = number of subjects at 5% significance level and 90% power.

BQ, Bournemouth Questionnaire; *CPQ,* Chronic Pain Questionnaire; *RODQ,* Revised Oswestry Disability Questionnaire; *MSPQ,* Modified Somatic Pain Questionnaire; *FABQ,* Fear Avoidance Beliefs Questionnaire; *PLC,* Pain Locus of Control questionnaire.

In contrast, Table 4 shows the correlations between total change scores of the BQ and total change scores of the external measures. As expected, the correlations were lower for the BQ and single aspect measures (PLC, MSPQ, Zung, and FABQ) and higher for the multiaspect measures (RODQ and CPG). All correlations were statistically significant (*P* <.02), indicating acceptable external longitudinal construct validity of total scores of the BQ.

Responsiveness

Table 5 shows the internal longitudinal construct validity or responsiveness of items in the questionnaire. The correlations between the item change scores and the corrected total change scores were all greater than 0.56. Because none of these correlations was obviously weaker than the others, it was inferred that all the items are responsive to clinically significant change and contribute significantly to the overall responsiveness of the BQ. On the basis of these data, none of the items is redundant to the BQ in terms of its ability to detect change over time.

On the basis of effect size, the results in Table 6 show the BQ to be as responsive as the RODQ and the CPG, necessitating a patient sample of a similar order of magnitude as these 2 established measures to detect clinical change at a statistically significant level. It should be noted that all the single trait measures (PLC, MSPQ, Zung, and FABQ) were less responsive in detecting clinical change than multiaspect measures (BQ, RODQ, and CPG) (Table 6). Moreover, the initial pretreatment scores (*n* = 55), in percentage terms, were similar between the RODQ and the BQ (43.7% ± 20.0% and 50.3& ± 18.8%, respectively) as were the posttreatment scores (22.3% ± 18.4% and 26.0% ± 20.1%, respectively).

DISCUSSION

Developing and testing a new questionnaire is a time-consuming task that should only be undertaken if a need exists. In this study, 3 separate phases of data collection were undertaken over a considerable time before complete testing of the instrument was possible. The pilot study and phase 1 involved lengthy reworking of the questions in the development stage so as to achieve some degree of face validity, both to patients and researchers. Even so, we might have gone further and interviewed a sample of patients to find out exactly what they understood by each question. Despite this, however, we believe the questionnaire consists of items that are clear and unambiguous and broadly understood by patients.

The need for a new questionnaire was identified by trawling the literature on back pain outcome measures and being faced with many long and cumbersome measures and very few, if any, multidimensional condition-specific measures suitable for use in documenting patient outcomes in a busy clinic practice setting. Although undoubtedly such instruments must exist, little evidence, as far as we are aware, of their psychometric testing exists. In a recent article by Deyo and coworkers,[35] the case for standardization of outcome measures in back pain research is forcibly made. Similar to this study, these authors distinguished outcome measures for use in clinical trials from those for use in other types of outcomes research. With reference to the latter, a "parsimonious" 6-item core set of outcome measures was proposed covering pain symptoms, daily activity disability, and satisfaction with treatment.[35] Although each of the items was extracted from an established measure, the psychometric properties of the 6-item set itself were not tested.[35]

Setting about selecting items for an outcome measure suitable for use in clinical practice, we were particularly mindful of the multidimensional nature of back pain on the basis of the biopsychosocial model[36] and the current shifts in outcome measures reflecting the illness, as opposed to disease, and care, as opposed to treatment, models.[9] Consequently, we included those dimensions we judged relevant to the model and that were shown in the literature at the time (and still are) to explain, at least in part, the back pain experience and its consequences. In addition, we used the criterion that all the dimensions included must be responsive to clinically significant change. As a result, the symptom of pain intensity, daily functional activity and social activity, the affective dimensions of anxiety and depression, and the cognitive/behavioral aspects of fear-avoidance beliefs and self-efficacy beliefs of pain control were included in the final instrument. The retention of all the items also maintained the multidimensional nature of the outcome measure. Pain intensity has been shown to decrease after treatment in numerous studies of back pain patients (eg, Meade et al[37]). Like pain intensity, many studies have shown significant changes in functional activity after treatment interventions in back pain patients (eg, Beurskens et al[38]). Similarly, distress levels,[27] fear-avoidance beliefs,[29] and pain locus of control[28] have all been shown to contribute significantly to the back pain experience and change after treatment. Finally, although these dimensions were singled out for inclusion in the questionnaire, it would be incorrect to interpret this as these being the only attributes of the back pain experience that change after treatment. In the trade-off for brevity of the questionnaire, these aspects were selected over others on the basis of review of the literature and our judgment of the important domains of the back pain model.

Having selected the item pool, we next turned our attention to item scaling, meaning the response options available to patients in answering the questions. Intuitively, increasing response options on a scale will increase item responsiveness and therefore the ability to detect clinically significant change. We, and others, have shown that an 11-point NRS is as responsive as a visual analog scale in detecting change in pain intensity and is easier for patients to complete.[39,40] We have also shown that asking patients about their pain levels over the previous week is more responsive compared with asking patients to report on their present pain levels.[39] For these reasons, all scales in the questionnaire were 11-point NRSs asking patients to report on their usual levels of the domains of interest over a 1-week period.

To ensure that none of the items is redundant in terms of contributing to the overall score of the questionnaire and that each is responsive to clinically significant change, item-corrected total correlations and item change score-corrected total change score correlations (respectively) were determined. The results of both these analyses determined that all the original items should be retained in the final version of the questionnaire.

Considerable confusion exists in the literature regarding which psychometric tests should be applied in the testing of new questionnaires and the terminology used. For example, in recent development and testing of new outcomes measures in the lumbar spine no consistency of either approach or terminology is apparent.[17-19,38] We adopted the criterion that only those tests applicable to the testing of evaluative measures should be used and consequently based the psychometric testing primarily on the framework advocated by Kirshner and Guyatt.[24] Other psychometric tests, of which there are many, were therefore purposely not included.

The results of the internal consistency analysis (Cronbach's alpha) showed that the questionnaire is a homogeneous instrument tapping different aspects of the same attribute (ie, back pain experience) and that as a result the items can be summed to produce a total overall score. Because the 7 dimensions each have a maximum score of 10, it may be more convenient to express the total score of the BQ as a percentage.

Because no "gold standard" exists, the external construct validity of the questionnaire was tested against established measures purporting to measure the same domains as those included in the questionnaire. The importance of testing longitudinal construct validity in an evaluative measure has been well argued by Kirshner and Guyatt,[24] and as a result we included this particular property and the more commonly investigated construct validity. In all cases the questionnaire robustly correlated with these external measures not only in terms of absolute scores (construct validity) but also in terms of the change scores over time (longitudinal construct validity).

In contrast to reliability and validity, it is surprising that the psychometric property of responsiveness is often not tested in evaluative measures. This is something of a paradox considering the ability to detect clinical change is an essential requirement of a measure of outcome.[30] In this study responsiveness of the individual items of the questionnaire in those patients who reported improvement in their condition was tested by correlating the change in item scores with the corrected total change scores of the questionnaire (also termed internal construct validity[38]). As has been discussed earlier, all items were sensitive to the improvement reported by patients. In terms of the overall responsiveness of the instrument and comparison with that of the external measures, the effect size was calculated. The questionnaire demonstrated a large effect size and, apart from that of the CPG, was the largest of any of the external measures used in this study, confirming the ability of the BQ to detect clinical change.

Before leaving the discussion on psychometric testing of the questionnaire, we should explain why the data collected in the 3 phases of the study were subjected to the analyses in the way they were. Obviously, only the data in phase 3 could be subjected to external validity and external responsiveness testing because it was only in this phase that external measures were used. Similarly only data in phase 2 could be subjected to test-retest reliability testing because only in this phase were the conditions for satisfactory test-retest met. However, data from both phases 2 and 3 could have been subjected to internal consistency and internal responsiveness testing. For clarity purposes for this article, it was decided to present the internal testing of the questionnaire with phase 2 data and the external testing of the questionnaire with phase 3 data. In all cases where data from both phases could be used, the analysis of both sets of data gave rise to the same interpretations and conclusions. It remains the case that development and testing of a questionnaire over a prolonged time inevitably leads to this multiphasic approach and the consequential overlapping of data sets.

The questionnaire has limitations. The patients recruited to the study were convenience samples both from the college teaching clinic and from chiropractors' field practices. As such, they may not be representative of the population of chiropractic patients as a whole, nor may they be representative of other ambulatory back pain patients. No attempt was made to distinguish between acute and chronic back pain patients. Also, the questionnaire was not developed for use in patients with other musculoskeletal disorders. Finally, the mean baseline scores of the pretreatment questionnaire, although comparable to other outcome measures, only registered approximately 50% of the scale, suggesting that further modifications may be possible to increase its responsiveness.

CONCLUSION

We have developed a multidimensional questionnaire for use in the routine documentation of outcomes in back pain patients attending chiropractic outpatient clinics. The questionnaire can be completed quickly and has been shown to be reliable, valid, and responsive. As such, it may be considered for use in other ambulatory back pain populations and in clinical trials of back pain treatments. We do not claim that the items in the questionnaire entirely cover the back pain experience or that they are necessarily the best available. However, we do believe that together they provide a reasonably comprehensive assessment that encompasses most of the important dimensions of the back pain model. The development of such an instrument makes possible studies directly comparing outcomes in practice-based settings with those in research trials, an important issue in the current moves to evidence-based practice.

ACKNOWLEDGMENTS

The data in phase 1 were collected by Dr Nigel Hunt and staff at the Salisbury Chiropractic Clinic, Salisbury, England. The AECC clinic reception staff collected the data used in phase 2. Phase 3 data were collected by a number of chiropractors in their practices, who at the time were enrolled on the MSc Clinical Chiropractic program. Our thanks go to all those involved.

REFERENCES

1. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical epidemiology: a basic science for clinical medicine. 2nd ed. London: Little, Brown and Co; 1991.
2. Sackett DL, Wennberg JE. Choosing the best research design for each question. BMJ 1997;315:1636.
3. Long AF, Dixon P. Monitoring outcomes in routine practice: defining appropriate measurement criteria. J Eval Clin Prac 1996;2:71-8.
4. Hoiriis KT, Owens EF, Pfleger B. Changes in general health status during upper cervical chiropractic care: a practice-based research project. Chiropractic Res J 1997;4:18-25.
5. Nyiendo J, Haas M, Hondras MA. Outcomes research in chiropractic: the state of the art and recommendations for the chiropractic research agenda. J Manipulative Physiol Ther 1997; 20:185-200.
6. Koes BW, Assendelft WJJ, van der Heijden GJ, Bouter LM. Spinal manipulation and mobilization for low back pain: an updated systematic review of randomized clinical trials. Spine 1996;21:2860-73.
7. Waddell G, Feder G, McIntosh A, Lewis M, Hutchinson A. Low back pain evidence review. London: Royal College of General Practitioners; 1996.
8. Clinical Standards Advisory Group of the Department of Health. Management Guidelines for Back Pain. London: HMSO; 1994.
9. Bolton JE. Future directions for outcomes research in back pain. Eur J Chiropractic 1997;45:57-64.
10. Merskey H. The definition of pain. Eur J Psychiatry 1991;6:153-9.
11. Verhaak PFM, Kerssens JJ, Dekker J, Sorbi MJ, Bensing JM. Prevalence of chronic bending pain disorder among adults: a review of the literature. Pain 1998;77:231-9.
12. Jensen MP, Karoly P. Self-report scales and procedures for assessing pain in adults. In: Turk D, Melzack R, editors. Handbook of pain assessment. London: Guildford Press; 1992. p. 135-51.
13. Melzack R. The McGill pain questionnaire: major properties and scoring methods. Pain 1975;1:277-99.
14. Fairbank JCT, Couper J, Davies JB, O'Brien JP. The Oswestry low back pain disability questionnaire. Physiotherapy 1980; 66:271-3.
15. Roland M, Morris R. A study of the natural history of back pain. Part 1. Development of a reliable and sensitive measure of disability in low back pain. Spine 1983;8:141-4.
16. Hudson-Cook N, Tomes-Nicholson K, Breen AC. A revised Oswestry disability questionnaire. In: Roland M, Jenner J. editors. Back pain. New approaches to rehabilitation and education. Manchester: Manchester University Press; 1989. p. 187-204.
17. Ruta DA, Garratt AM, Wardlaw D, Russell IT. Developing a valid and reliable measure of health outcome for patients with low back pain. Spine 1994;19:1887-96.
18. Kopec JA, Esdaile JM, Abrahamowicz M, Wood-Dauphinee S, Lamping DL, Williams JI. The Quebec back disability scale. Spine 1995;20:341-52.
19. Daltroy LH, Cats-Baril WL, Katz JN, Fossel AH, Liang MH. The North American Spine Society lumbar spine outcome assessment instrument. Reliability and validity tests. Spine 1996;21:741-9.
20. Bolton JE. Evaluation of treatment in back pain patients: clinical outcome measures. Eur J Chiropractic 1994;42:29-40.
21. Bolton JE. Evaluation of the psychosocial profile of the chronic back pain patient. In: Stude D, editor. A professional's guide to spinal rehabilitation. Stamford (CT): Appleton & Lange. In press 1999.
22. Bergner MB, Bobbitt RA, Carter WB, Gilson BS. The SIP: development and final revision of a health status measure. Med Care 1981;19:787-805.
23. Brazier JE, Harper R, Jones NMB, et al. Validating the SF-36 health survey questionnaire: new outcome measure for primary care. BMJ 1992;305:160-4.

24. Kirshner B, Guyatt G. A methodological framework for assessing health indices. J Chron Dis 1985;38:27-36.
25. Streiner DL, Norman GR. Health measurement scales. 2nd ed. Oxford: Oxford University Press; 1995.
26. Smith BH, Penny KI, Purves AM, et al. The Chronic Pain Grade questionnaire: validation and reliability in postal research. Pain 1997;71:141-7.
27. Main CJ, Wood PLR, Hollis S, Spanswick CC, Waddell G. The distress and risk assessment method. A simple patient classification to identify distress and evaluate the risk of poor outcome. Spine 1992;17:42-52.
28. Toomey TC, Seville JL, Mann JD. The pain locus of control scale: relationship to pain description, self-control skills and psychological symptoms. The Pain Clinic 1995;8:315-22.
29. Waddell G, Newton M, Henderson I, Somerville D, Main CJ. A fear-avoidance beliefs questionnaire (FABQ) and the role of fear-avoidance beliefs in chronic low back pain and disability. Pain 1993;52:157-68.
30. Bolton JE. On the responsiveness of evaluative measures. Eur J Chiropract 1997;45:5-8.
31. Stratford P, Solomon P, Binkley J, Finch E, Gill C. Sensitivity of Sickness Impact Profile items to measure change over time in a low-back pain patient group. Spine 1993;18:1723-7.
32. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. Controlled Clin Trials 1991;12:142S-58S.
33. Kaziz LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. Med Care 1989;27:S178-S89.
34. Machin D, Campbell M, Fayers P, Pinol A. Sample size tables for clinical studies. 2nd ed. Cambridge (MA): Blackwell Science; 1997.
35. Deyo RA, Battie M, Beurskens AJHM, et al. Outcome measures for low back pain research. A proposal for standardized use. Spine 1998;23:2003-13.
36. Waddell G. A new clinical model for the treatment of low back pain. Spine 1987;12:632-44.
37. Meade TW, Dyer S, Browne W, Townsend J, Frank AO. Low back pain of mechanical origin: randomized comparison of chiropractic and hospital outpatient treatment. BMJ 1990; 300:1431-7.
38. Beurskens AJ, deVet HC, Koke AJ, van der Heijden GJ, Knipschild PG. Measuring the functional status of patients with low back pain. Spine 1995;20:1017-28.
39. Bolton JE, Wilkinson RC. Responsiveness of pain scales: a comparison of three pain intensity measures in chiropractic patients. J Manipulative Physiol Ther 1998;21:1-7.
40. Jensen MP, Miller L, Fischer LD. Assessment of pain during medical procedures: a comparison of three scales. Clin J Pain 1998;14:343-49.

APPENDIX 1

Global Dimensions of the Bournemouth Questionnaire

The following scales have been designed to find out about your back pain and how it is affecting you. Please answer ALL the scales by circling ONE number on EACH scale that best describes how you feel:

1. Over the past week, on average, how would you rate your back pain?

No pain — Worst pain possible
0 1 2 3 4 5 6 7 8 9 10

2. Over the past week, how much has your back pain interfered with your daily activities (housework, washing, dressing, walking, climbing stairs, getting in/out of bed/chair)?

No interference — Unable to carry out activity
0 1 2 3 4 5 6 7 8 9 10

3. Over the past week, how much has your back pain interfered with your ability to take part in recreational, social, and family activities?

No interference — Unable to carry out activity
0 1 2 3 4 5 6 7 8 9 10

4. Over the past week, how anxious (tense, uptight, irritable, difficulty in concentrating/relaxing) have you been feeling?

Not at all anxious — Extremely anxious
0 1 2 3 4 5 6 7 8 9 10

5. Over the past week, how depressed (down-in-the-dumps, sad, in low spirits, pessimistic, unhappy) have you been feeling?

Not at all depressed — Extremely depressed
0 1 2 3 4 5 6 7 8 9 10

6. Over the past week, how have you felt your work (both inside and outside the home) has affected (or would affect) your back pain?

Have made it no worse — Have made it much worse
0 1 2 3 4 5 6 7 8 9 10

7. Over the past week, how much have you been able to control (reduce/help) your back pain on your own?

Completely control it — No control whatsoever
0 1 2 3 4 5 6 7 8 9 10