Responsiveness of Shuttle Walking Test Compared With Disease Specific and Generic Outcome Measures in Patients with Chronic Back Pain

O Rivero-Arias[1], HE Campbell[1], K Johnston[2] and JCT Fairbank[3] on behalf of the Spine Stabilisation Trial Group

[1] Health Economics Research Centre, Institute of Health Sciences, University of Oxford, Oxford, UK.

[2] Economics and Statistics Division, Scottish Executive, Edinburgh, UK

[3] Nuffield Orthopaedic Centre, Oxford, UK

Source of support: Medical Research Council

Abstract

Study Design. A prospective cohort of 202 patients with chronic back pain randomised to either surgery stabilisation of the spine or intensive rehabilitation followed up for 24 months.

Objectives. To determine the responsiveness of the Shuttle Walking test (SWT) (a dimension specific outcome measure measuring metres walked) compared to a disease specific measure Oswestry Disability Index and two generic outcome measures (SF-36 and EQ-5D).

Summary of Background Data. Although there are a number of studies assessing the responsiveness of outcome measure in back pain, no study has explored the responsiveness of the shuttle walking test relative to other outcomes measures for patients with chronic back pain.

Methods. The shuttle walking test was administered at a rehabilitation clinic. The Oswestry disability index, SF36 and EQ-5D were assessed by questionnaire at baseline and 24 months. Responsiveness was assessed using conventional measures such as effect size, standardised response mean (SRM) and receiver operating (ROC) curves.

Results. Mean figures of each instrument suggest an improvement in outcomes over time. The physical component of the SF-36 yielded the highest effect size (1.65) and the ODI the highest SRM (1.23) in the improved group. The lowest effect size and SRM for the improved group was recorded on the mental component of the SF-36 at -0.20 and -0.18 respectively. The greater responsiveness in the ROC curves among the instruments was achieved by the ODI and the SF-36 physical component.

Discussion. The results presented in this paper appear to demonstrate that the ODI, EQ-5D and the SF-36 physical component are more sensitive to change in patients with chronic low back pain than the SWT. The large sample size and the consistency of the different methods across the improved and non-improved groups support the results achieved.

Introduction

A large number of patient based outcome measures are now available for research. These can generally be classified into three types: dimension specific measures (measuring one dimension of health related quality of life); disease specific measures (measuring a number of dimensions of health related quality of life relevant to a particular disease); and generic measures (measuring a number of dimensions relevant to a wide range of conditions). The latter type of measure can be further divided into those that are health profiles (where the items in the profile are scores and summed to produce a score for each dimension), such as SF36, and those used to generate utility measures (where the measure produces a single overall score (or utility) from the dimensional measures), such as the EuroQol EQ-5D.

In many studies generic instruments have been included alongside disease specific instruments, sometimes purely for comparative purposes, more recently to facilitate the measurement of quality adjusted life years (QALYs) when cost-effectiveness is an endpoint secondary to clinical effectiveness. Clinical trials of interventions for low back pain, which have increased in number over the last decade, provide a good example of where this is the case (Koes, B.W. 1995). Low back pain although not a life threatening condition, can generate great discomfort among individuals of working age (Clinical Standard Advisory Group 1994). The annual incidence has been estimated to be around 5% and the annual prevalence lies between 15% and 40% in most Western countries (Frymoyer, J.W et al. 1991). In addition, the direct costs associated with the disability are high and were estimated to be around £1632 million in the UK in 1998 (Maniadakis and Gray 2000). Studies exploring the outcomes of different treatments and interventions for back pain patients often use disease specific outcome measures (Bombardier, C. 2000), such as the Oswestry Disability Index (ODI) (Fairbank and Pynsent 2000) or the Roland-Morris Questionnaire (Roland and Morris 1983) in addition to a generic measure such as the SF36 or EQ-5D.

The Shuttle Walking Test (SWT) (Singh et al. 1994) is a measure of mobility and is rarely used as an outcome measure in back pain patients. The purpose of this paper is to assess the responsiveness of the SWT in a population of chronic back pain patients

against one disease specific (ODI) and two generic outcome measures (SF-36, EQ-5D). Responsiveness addresses sensitivity to change, that is, whether a health status questionnaire can detect important changes over time (Fitzpatrick et al. 1998). Past research in the area of back pain has compared the responsiveness of several disease specific and generic outcome measures (Beurskens et al. 1996, Taylor, S. et al. 2001, Frost H. et al. 2000) however there are no studies in this area comparing the performance of the SWT against such disease specific and generic instruments. This paper presents such an analysis, using the traditional measures of responsiveness of effect size, standardised response mean and receiver operating characteristics curves (Deyo et al. 1991).

Materials and methods

Data

Two hundred and two patients (202), recruited to the Medical Research Council Spine Stabilisation trial (SST), between June 1996 and June 2000, were included in this analysis. The aim of this pragmatic randomised controlled trial was to determine whether surgical treatment of patients with chronic back pain by spinal stabilisation is more or less cost-effective than non-operative treatments in achieving a worthwhile relief of symptoms.

Patients between 18 and 55 years of age, who were being considered for first time surgical stabilisation of the spine, and who had a history of chronic back pain exceeding 12 months, were assessed by surgeons to determine their eligibility for the trial. Patients for whom surgeons were uncertain as to whether spine stabilisation or intensive rehabilitation would prove more beneficial were considered suitable for randomisation.

The health status of trial patients was measured at baseline and 24 months. The SWT was administered at a rehabilitation clinic. The ODI, SF-36 and EQ-5D were assessed by postal questionnaire.

Description of the instruments

Shuttle walking test

The SWT was originally developed for the assessment of respiratory function in patients with obstructive pulmonary disease (Singh et al. 1994) but has subsequently been used to assess patients with chronic heart failure (Keell et al. 1998), pacemaker implants (Payne and Skehan, 1996) and brain injuries (Vitale et al. 1997). SWT has been shown to correlate well with measures of oxygen uptake (VO_2max) in patients with heart disease (Keell et al. 1998; Lewis et al. 2001).

The SWT is a dimension specific measure of mobility as measured by the number of metres walked in experimental conditions, usually administered in a clinic. The SWT requires patients to walk up and down a 10 metre course at speeds dictated by signals from an audio tape recorder. The walking speed is increased by a small increment for each additional minute. For example, in the first minute the patient is required to walk up and down the 10 metre course three times (amounting to a distance of 30 metres). In the second minute, the patient is required to walk up and down the 10 metre course four times (amounting to a distance of 40 metres) and so on. The test is stopped either by the patient stopping because of increased fatigue and failure to complete a 10 metre course or by the operator when the patient fails to complete the required distance within the minute. The maximum number of minutes attempted is 12 (minimum one minute) and the corresponding maximum distance achieved is 1020 metres (minimum 10 metres) The higher the number of metres walked, the better the patient's mobility.

Oswestry Disability Index

The ODI is a disease specific outcome measure and is one of the principal outcome measures used in the management of spinal disorders (Fairbank and Pynsent 2000). The ODI has also been used as an outcome measure in clinical trials (Anderson et al 1999). The ODI has ten dimensions: pain intensity, personal care (washing dressing etc.), lifting, walking, sitting, standing, sleeping, sex life (if applicable), social life and travelling. Each dimension has 6 levels, with a score of zero allocated to the best level and a score of five allocated to the worst level. The maximum score of the ODI is therefore 50 but the ODI is converted to a percentage with a consequent maximum

of 100%. A clinically meaningful change in the ODI of 4 percentage points is being used in the SST (Fairbank and Pysnent 2000). The ODI has been found to correlate well with the SF36 (Grevitt et al. 1997). Although the ODI has been validated, additional information on its responsiveness is required (Fairbank and Pysnent 2000). Since a lower score on the ODI implies better health status, for the purposes of the comparisons between measures, the ODI score is inverted so that a higher score indicates better health status.

SF36

The SF36 is a generic health profile instrument containing 36 items reducing to summary scores of eight dimensions (general health perception, physical functioning, role (physical), role (emotional), pain, social functioning, mental health, energy/vitality) (Jenkinson et al. 1997). For each of the eight dimensions, item scores are coded, summed and transformed to a scale of 0 (worst possible health state) to 100 (best). The coding for the UK SF36 was used (Jenkinson et al. 1997). The SF36 was further combined into the two summary scores: physical component score (PCS) and the mental health component score (MCS) (Jenkinson et al. 1997).

EQ-5D

The EQ-5D is a generic utility measure of health status (EuroQol Group 1990). It has five questions covering five dimensions of mobility, self-care, usual activity, pain/discomfort, and anxiety/depression. Each question has three response categories: level 1 "no problems", level 2 "some problems", level 3 (inability or extreme problems". The responses to these items therefore combine to give a descriptive health state classification with five digits, e.g. 11222. The EQ-5D yields a total of 243 possible health states (i.e. 3^5). A set of valuations for each health state is available (Dolan et al. 1995) and is derived from a sample of 2,997 individuals in England, Scotland and Wales. The EQ-5D measure was developed specifically for use in economic evaluation and the minimum and maximum tariff scores are –0.594 and 1 respectively.

Measures of responsiveness

Effect size, standardised response mean (SRM) and receiver operating characteristic curves were used to assess instrument responsiveness between baseline and 24 months. A description of each technique is presented below.

Effect size

A common form of expression of responsiveness is the effect size. It measures the size of change on a measure that occurs to a group between assessments compared with the variability of scores of that measure (Fitzpatrick et al. 1998). Effect sizes are calculated as the difference between mean scores at assessments divided by the standard deviation of the baseline score. This transforms the score change into a standard unit of measurement, which can then be compared with scores changes of other instruments (which may be in different units) (Deyo et al. 1991). Effect sizes can be translated into benchmarks for assessing the relative size of change, an effect size of 0.2 being considered small, 0.5 as medium and 0.8 or greater as large (Cohen 1978). An effect size of 1 is equivalent to a change of one standard deviation in the sample.

Standardised response mean (SRM)

Another approach for measuring responsiveness is the SRM. Its definition is similar to the effect size, the method of calculation differing only in that the difference between mean scores at follow-up points is divided by the standard deviation of the change in scores between groups, rather than the standard deviation of the baseline scores (Liang et al. 1990). As with the effect size, the method of calculating the SRM facilitates the direct comparison of instruments expressing outcomes using different units. In addition, Cohen's rule-of-thumb for interpreting effect sizes can also be applied.

For this analysis, comparisons between instruments using both effect size and the SRM techniques were carried out for two groups of patients, those who perceived their health as having improved during the study duration and those who considered their health to have deteriorated. Such information was gleaned from a question on the (24-month) SF36 questionnaire, which asks respondents to rate their current

general health compared to one year ago. Answers are given on a five point scale where 1 = much worse now than one year ago, 2 = somewhat worse now than one year ago, 3 = about the same, 4 = somewhat better now than one year ago, 5 = much better now than one year ago. Respondents whose answers scored 4 or 5 were classified as having a health improvement, those scoring three or less were considered to have a health status unchanged or deteriorated.

Receiver-operating characteristic curves

Deyo et al. (1991) suggested the idea of identifying whether a health outcome distinguishes changes over time the same way a diagnostic test identifies patients with a particular characteristic of interest. Using this approach, the change in health status scores is examined in terms of the sensitivity of change scores produced by an instrument (proportion of true changes detected) and 1-specificity of change scores (proportions of individuals who are detected as changing but are truly stable). The sensitivity and (1 minus specificity) are then plotted against each other to give a receiver operating characteristic (ROC) curve. The most responsive instrument would have a plot where the true positive rate sharply increases whilst the false positive rate remains low. The greater the total area under a curve from all cut-off points, the greater the instruments responsiveness (Deyo et al. 1991). Consequently, an area of 0.5 indicates no accuracy in detecting change and an area of 1.0 indicates perfect accuracy (Deyo et al. 1991). Confidence intervals for the different areas can be computed to quantify the uncertainty around the point estimates.

For the purposes of calculating the sensitivity and specificity of each instrument, and for constructing the ROC curve, an external 'gold standard' evaluation of whether a change in health status has actually occurred is required. Whether or not a patient had actually experienced a true improvement in health was again determined using answers from the self-perception of health question on the 24-month SF36. Respondents with answers scoring 4 or 5 were classified as having a true health improvement, those scoring three or less were considered to have experienced no health improvement.

Results

Patient characteristics

Analysis was carried out using data obtained from 202 trial patients. The mean age (standard deviation) of these patients was 40 (8.7) years (range 19-55 years) and 110 (54.4%) of the sample were women. The mean duration (standard deviation) of low back pain (in years) was 7.6 (6.7) with a range from 1 to 35 years. Finally 109 (54%) of the patients were in paid employment.

Table 1 shows how patients rated their current health compared to one year ago on the 24-month SF36 questionnaire. Eighty-eight (44%) patients reported their health as being 'somewhat better now than one year ago', or 'much better now than one year ago', and were classified as having a health improvement. Patients were categorised as non-improvers if they reported no change or a deterioration in health status. One hundred and fourteen (56%) patients fell into this category.

Table 2 shows baseline and 24-month mean scores and standard deviations for each instrument by patient group (i.e. improved and non-improved patients). All of the instruments with the exception of the MCS (which remained relatively stable), exhibit large score increases over time in the improved group. The SWT score for example increases from 279.43 to 413.07. Improvements although much smaller, are also detected by each instrument for patients classified as not having experienced an improvement in health.

Effect Size and SRM

The size of the mean score change on each instrument differed between the improved and non-improved groups (table 3). The effect size and the standardised response mean were higher in the improved than in the non-improved group for all the instruments (again with the exception of the MCS). The PCS yielded the highest effect size (1.65) and the ODI the highest SRM (1.23) in the improved group. The lowest effect size and SRM for the group was recorded on the MCS at -0.20 and -0.18 respectively. In the non-improved group the highest effect size was achieved by the

PCS (0.49) and the SWT instrument (0.44) yielded the highest SRM. The MCS had the lowest effect size (0.09) and SRM (0.09) in the non-improved group.

Setting aside the mental component of the SF-36, table 3 shows that in the improved group, when compared with all remaining instruments, the SWT was the least responsive when assessed using both the effect size and the SRM. Despite this finding, the level of change which the SWT was able to detect would according to Cohen's rule of thumb, still be classified as medium to large. For the non-improved group, the SWT was at least as responsive according to the effect size, as the EuroQol EQ-5D. In contrast the SRM shows the SWT to be the most responsive instrument for this group.

Receiver-operating characteristic curve

Figure 1 shows the ROC curves for the different instruments constructed using the external gold standard of patient-perceived change in health. The most responsive instrument is one producing the curve closest to the upper left hand corner of the box, i.e. higher likelihood of finding a significant change when the change has really occurred (sensitivity) and lower likelihood of finding a significant change when the change has not really occurred (1-specificity). The area under the ROC curve indicates the probability of identifying correctly improved patients from randomly selected pairs of improved and unimproved patients (Deyo and Centor 1985).

Figure 1 shows the curves for the ODI and the PCS to be closest to the upper left hand corner of the box than the curves for the others instruments. Table 4 presents the area under the ROC curve for each of the four instruments that demonstrated some ability for detecting change (The MCS had no accuracy (area value 0.46 95% CI. 0.34 to 0.51) and was therefore excluded from the figure). As suggested by the position of their curves, these area values were greatest for the ODI (0.77, with 95% CI 0.70 to 0.83) and the PCS (0.74,95% CI 0.67 to 0.81). The area under the ROC curve calculated from the EQ-5D was slightly smaller (0.70 with 95% CI 0.62 to 0.77). Of the four instruments, the SWT had the smallest area (0.62 95% CI 0.54 to 0.70).

The cut-off points that best discriminate between improved and non-improved are shown in table 4. Scores changes with the best cut-off points were: 55-56 meters on the SWT, 8-10 points on the inverted 100 points ODI scale, 0.06- 0.1 on the EQ-5D on a scale from 0 to 1 and 14-17 points on the 100 points physical component scale of the SF-36.

Discussion

This paper aimed to compare the responsiveness of a dimension specific end point (the SWT) with that of a disease specific (the ODI) and two generic (the SF36 and EQ-5D) patient based outcome measures in the area of chronic back pain. Measurements from each instrument were available at baseline and 24 months for two hundred and two patients recruited to the MRC Spine Stabilisation Trial. This constitutes a fairly large sample when compared to other studies assessing the responsiveness of instruments for use in low back pain or orthopaedic evaluation (Liang et al. 1990, Beurskens et al. 1996).

Given the lack of consensus in the published literature as to the most appropriate technique for measuring responsiveness, we used three different methods; effect size, standardised response means (SRM) and receiver operating characteristic (ROC) curves. The external 'gold standard' criterion used to determine whether an improvement in health had actually occurred between assessment points, was a patient's self-perceived health improvement as measured at 24 months using the SF-36 questionnaire.

Results appear to indicate a broad level of agreement between each of the three measures of responsiveness. Each classifies the SWT for example as being the least responsive when considered alongside the ODI, the EQ-5D, and the PCS. Possible explanations for such a finding are that the SWT measures functional status purely in terms of mobility. Although walking is an important functional activity, there are other dimensions that patients may consider necessary for relieving their low back pain symptoms. These other categories may well be captured on the other instruments.

The ODI, for instance, includes ten dimensions of which mobility is just one. Similary, of the EQ-5Ds five dimensions, only one relates to mobility. Despite this, the SWT has still demonstrated an ability to detect change in a patient's facility to walk, an important daily function.

With the exception of the MCS, the results suggest that all of the instruments were able to discriminate between patients classified as improved and non-improved. Effect sizes and SRMs were moderate to large for these four instruments in the improved group, however table 3 also reveals positive effect sizes and SRMs that are small to moderate in magnitude in the non-improved group. This indicates that some patients, who were classified as non-improvers based upon their self-perception of health, did register improvements on the other instruments. This focuses attention towards the question used to elicit patients' views about their own health and suggests that despite reporting improvements on specific questionnaires, some patients did not perceive these improvements to have be large enough to rate their health as being at least 'somewhat better now than one year ago'. Given the absence of an option 'slightly better now than one year ago', patients may have simply classified themselves as 'about the same'.

Small to moderate positive effect sizes in patients classified as non-improvers using an external gold standard have been observed elsewhere. In the area of low back pain, Beurskens et al. (1996) demonstrated quite clearly how a moderate positive effect size observed on an instrument for patients classified as non-improvers, was reduced when patients reporting slight improvements (initially labelled non-improvers) were reclassified according to the external criteria, as improvers.

Such scenarios highlight the dependency of responsiveness evaluation on external criteria for judging health improvements. In this study reliance is placed upon a patient's own judgement as to whether or not an improvement in health has been experienced. Use of other external criteria may well be associated with different findings.

The results presented in this paper appear to demonstrate that the ODI, EQ-5D and the SF-36 physical component are more sensitive to change in patients with chronic low

back pain than the SWT. The large sample size and the consistency of the different methods across the improved and non-improved groups support the results achieved.

References

Beurskens AJHM, De Vet HCW, Koke AJA. Responsiveness of functional status in low back pain: A comparison of different instruments. *PAIN* 1996;**65**:71-6.

Bombardier C. Outcome assessments in the evaluation of treatment of spinal disorders: summary and general recommendations. *Spine* 2000;**25**:3100-3.

Clinical Standards Advisory Group. Epidemiology Review: *The Epidemiology and Cost of Back Pain*. 1994. London, HMSO.

Cohen J. Statistical power analysis for the behavioural sciences. New York: Academic Press, 1978.

Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. *Controlled Clinical Trials* 1991;**12**:142S-58S.

Dolan P, Gudex C, Kind P, Williams A. A social tariff for EuroQol: Results from a general population survey. York. United Kingdom: University of York. *Centre for Health Economics*, 1995. Discussion article 138.

EuroQol Group. EuroQol - A new facility for the measurement of health-related quality of life. *Health Policy* 1990;**16**:199-208.

Fairbank JC,.Pynsent PB. The Oswestry Disability Index. *Spine* 2000;**25; discussion 29522**:2940-52.

Fitzpatrick, R., Davey, C., Buxton, M. J., and Jones, D. R. Evaluating patient-based outcome measures for use in clinical trials. Health Technology Assessment. 2(14), i-74. 1998.

Frost H, Lamb SE, Shackleton CH. A functional restoration programme for chronic low back pain. *Physiotherapy.* 2000;**86**:285-93.

Frymoyer JW,.Cats-Baril WL. An overview of the incidences and costs of low back pain. *Orthopedic Clinics of North America.* 1991;**22**:263-71.

Grevitt M, Khazim R, Webb J, Mulholland R, Shepperd J. The short form-36 health survey questionnaire in spine surgery. *J-Bone-Joint-Surg-Br.* 1997;**79**:48-52.

Jenkinson C, Layte R, Lawrence K. Development and testing of the Medical Outcomes Study 36-Item Short Form Health Survey summary scale scores in the United Kingdom. Results from a large-scale survey and a clinical trial. *Med.Care* 1997;**35**:410-6.

Keell SD, Chambers JS, Francis DP, Edwards DF, Stables RH. Shuttle-walk test to assess chronic heart failure. *LANCET* 1998;**352**:705.

Koes BW, Bouter LM, van der Heijden GJ. Methodological quality of randomized clinical trials on treatment efficacy in low back pain. *Spine* 1995;**20**:228-35.

Lewis ME, Newall C, Townend JN, Hill SL, Bonser RS. Incremental shuttle walk test in the assessment of patients for heart transplantation. *Heart* 2001;**86**:183-7.

Liang MH, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopedic evaluation. *Med.Care* 1990;**28**:632-422.

Maniadakis N,.Gray A. The economic burden of back pain in the UK. *PAIN* 2000;**84**:95-103.

Norman GR, Streiner DL. Health Measurement Scales: a practical guide to their development and use. Oxford University Press, 1995.

Payne GE,.Skehan JD. Shuttle walking test: a new approach for evaluating patients with pacemakers. *Heart* 1996; **75**:414-8.

Roland M,.Morris R. A study of the natural history of back pain. Part I: Development of a reliable and sensitive measure of disability in low-back pain. *Spine* 19;**8**:141-4.

Singh SJ, Morgan MD, Hardman AE, Rowe C, Bardsley PA. Comparison of oxygen uptake during a conventional treadmill test and the shuttle walking test in chronic airflow limitation. *Eur.Respir.J.* 1994;**7**:2016-20.

Taylor S, Frost H, Taylor A, Barker K. Reliability and responsiveness of the shuttle walking test in patients with chronic low back pain. *Physiother-Res-Int* 2001;**6**:170-8.

Taylor SJ, Taylor AE, Foy MA, Fogg AJ. Responsiveness of common outcome measures for patients with low back pain. *Spine* 1999;**24**:1805-12.

Vitale AE, Jankowski LW, Sullivan SJ. Reliability of a walk/run test to estimate aerobic capacity in a brain-injured population. *BRAIN.INJ.* 1997;**11**:67-76.

Table 1: Patients health status compared to one year ago (n=202)

Answer	Frequency	(%)
Much worse now than one year	11	(6)
Somewhat worse now than one year ago	33	(16)
About the same	70	(35)
Somewhat better now than one year ago	57	(28)
Much better now than one year ago	31	(15)

Table 2: Mean scores (SD) of the instruments at baseline and 24 months follow-up points in the improved and non-improved groups

Instrument	Baseline				24-months			
	Improved (n=88)		Non-improved (n=114)		Improved (n=88)		Non-improved (n=114)	
SWT	279.43	(190.77)	213.95	(188.52)	413.07	(216.93)	276.49	(219.59)
ODI *	58.18	(13.39)	52.45	(14.61)	78.05	(14.44)	57.39	(18.53)
EQ-5D	0.43	(0.30)	0.34	(0.32)	0.76	(0.19)	0.45	(0.35)
PCS	27.64	(18.75)	23.10	(18.24)	58.57	(26.53)	32.07	(25.61)
MCS	62.69	(17.39)	61.28	(17.38)	59.17	(14.65)	62.92	(17.93)

* The ODI score is inverted so that a higher score indicates better health status

Table 3: Mean score changes, effect sizes and SRMs in the improved (n=88) and non-improved (n=114) groups.

Instrument	Mean Score Change	SD at baseline	Effect Size	SD of the mean change	SRM
SWT					
Improved	133.64	(190.77)	0.70	(186.56)	0.72
Non-improved	62.54	(188.52)	0.33	(142.72)	0.44
ODI*					
Improved	19.87	(13.39)	1.48	(16.16)	1.23
Non-improved	4.94	(14.61)	0.34	(12.92)	0.38
EQ-5D					
Improved	0.33	(0.30)	1.09	(0.32)	1.02
Non-improved	0.11	(0.32)	0.33	(0.31)	0.35
PCS					
Improved	30.93	(18.75)	1.65	(25.97)	1.19
Non-improved	8.97	(18.24)	0.49	(22.02)	0.41
MCS					
Improved	-3.52	(17.39)	-0.20	(20.11)	-0.18
Non-improved	1.64	(17.38)	0.09	(18.71)	0.09

* The ODI score is inverted so that a higher score indicates better health status

Figure 1: ROC curve of the scores changes for the SWT, ODI, EQ-5D and PCS

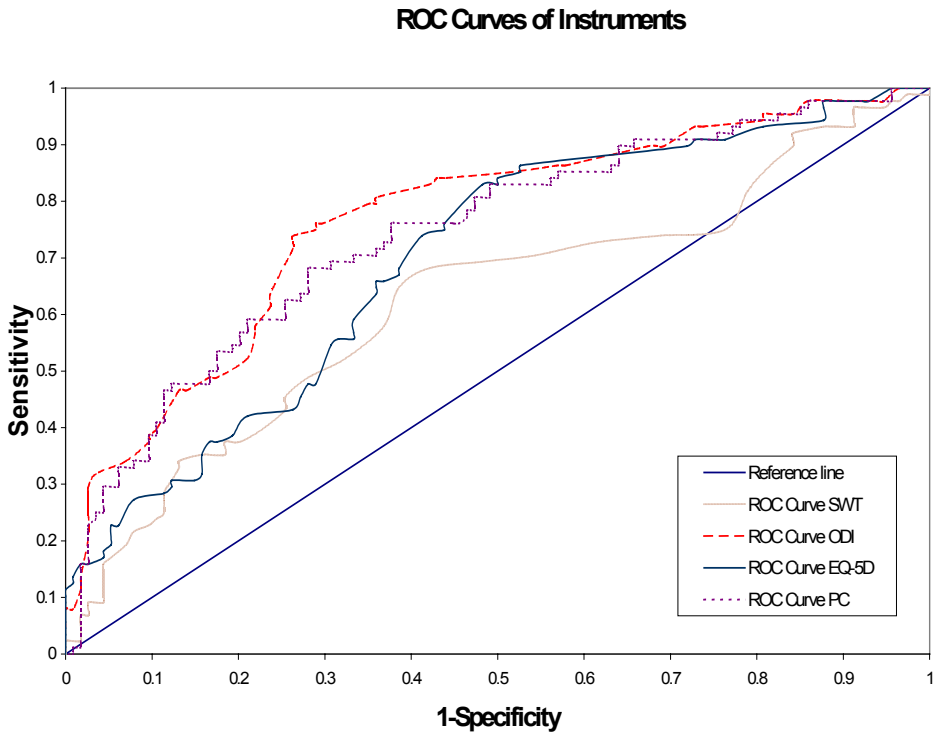

Table 4: Area under the ROC curve for each instrument

Instruments	Area	95% Confidence Interval	Optimal cut-off on instrument
SWT	0.62	(0.54 to 0.70)	55-65
ODI*	0.77	(0.70 to 0.83)	8-10
EQ-5D	0.70	(0.62 to 0.77)	0.06-0.1
PCS	0.74	(0.67 to 0.81)	14-17
MCS	0.43	(0.34 to 0.51)	NS

* The ODI score is inverted so that a higher score indicates better health status
NS Non significant