International Journal of Technology Assessment in Health Care, **17:4** (2001), 479–487.

REPRODUCIBILITY AND RESPONSIVENESS OF EVALUATIVE OUTCOME MEASURES

Theoretical Considerations Illustrated by an Empirical Example

Henrica C. W. de Vet
Lex M. Bouter
P. Dick Bezemer
Vrije Universiteit Medical Centre

Anna J. H. M. Beurskens
Maastricht University

Abstract

Objective: This article outlines basic methods for quantifying reproducibility and responsiveness of evaluative outcome measures.
Methods: The background noise in stable patients provides the desired information to quantify the reproducibility. From this, the smallest real difference (SRD) for longitudinal differences can be derived. We propose to use the SRD to define responsiveness: from all patients who change according to an external criterion, we take the percentage that changes at least SRD on the outcome measure. A more complete picture of the responsiveness of the outcome measure arises when the receiver operating characteristic (ROC) is constructed, corresponding to all possible cut-off change scores. The proposed methods are illustrated with an empirical example.
Results: In the illustration the methods appeared to be very useful and complemented each other. We could evaluate whether the observed change score was larger than that expected due to chance. With the methods it was possible to evaluate both the ability of an instrument to detect change if there is a real change in health status (sensitivity to change) and the ability to detect absence of change if there is no real change (specificity to change).
Conclusion: We presented the use of SRDs and ROC curves for quantifying reproducibility and responsiveness. We started with the basic notions and arrived at methods that are both understandable and useful.

Keywords: Reproducibility, Responsiveness, Outcome measures, Clinimetrics, ROC curve

The methodologic quality of instruments measuring health status has received considerable attention over the last few years (5;12;13;21;27). Evaluative outcome measures are used to assess the magnitude of longitudinal change in patients, for example, in clinical trials or in

daily practice. The requirements for evaluative outcome measures are high reproducibility and high responsiveness (9;15).

In this article we try to develop a single and tenable framework starting from the most basic methods for quantifying reproducibility and responsiveness. The proposed methods will be illustrated by an example from a randomized controlled trial on traction for nonspecific low back pain (3). Reproducibility is the extent to which the same results are obtained on repeated administrations of the same instrument when no real change in health status has occurred. For an evaluative outcome measure, it is important that repeated measurements for stable individuals remain constant over time (15;21). Lack of reproducibility can be the consequence of random measurement error (27) and real within-person variance (9;21). Both components together lead to measured fluctuations in health status that occur in the absence of real change over time: the "background noise."

Responsiveness is the extent to which different results are obtained on repeated administrations of the same instrument when a real change in health status has occurred (9;17;19;28). To assess the responsiveness of an evaluative outcome measure, often an external criterion is used to define whether a patient has changed (26). The use of an implicit external criterion also has been proposed (23). In our example we use an explicit external criterion derived from the same study to determine whether a patient's back complaints have deteriorated, showed no change, or have improved (Figure 1). The external criterion determines the minimum change that is considered to be clinically relevant (14;20;32;33). With the use of an external criterion we are able to identify patients who change equal to or more than this minimal amount of clinically relevant change. In order to facilitate the interpretation of treatment effects, Guyatt et al. (16) have proposed to use fixed criteria to define clinically important differences.

METHODS

Reproducibility

For an evaluative outcome measure, it is important that the reproducibility in stable patients is as high as possible. The most basic statistic for quantifying reproducibility is the standard deviation of a single measurement (SD_{single}) taken at any moment for a stable patient. SD_{single} contains random measurement error as well as within-person variance. To evaluate the reproducibility within subjects, we are interested in the subject variability of the individual changes in scores over time. If a person takes the same test on two occasions, the standard deviation of the *difference* between the two scores can be computed directly. This is the relevant statistic when evaluating changes in an individual, abbreviated in this article as

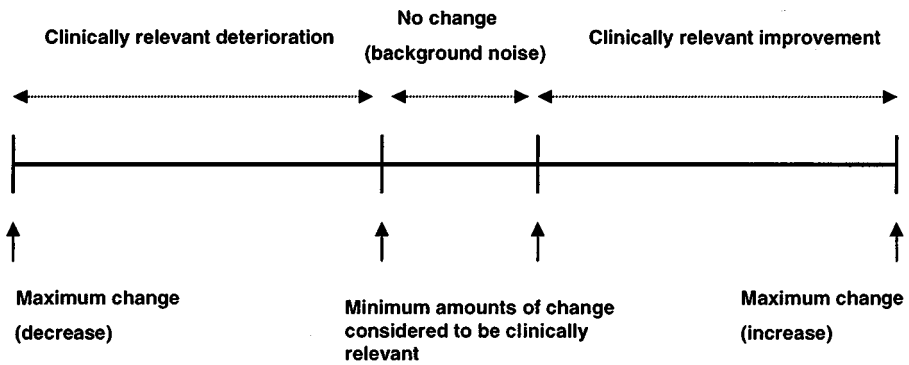

Figure 1. Available options for change over time in disease severity.

https://doi.org/10.1017/S0266462301107038 Published online by Cambridge University Press

$SD_{change\ ind}$. Dividing by $\sqrt{2}$ produces the basic statistic SD_{single}, which can also be assessed by giving the same test repeatedly (27).

To evaluate whether the observed change is larger than expected due to background noise alone, the relevant statistic ($SD_{change\ ind}$) should be multiplied by 1.96. Observed change of at least this size implies a real change with 95% confidence. We propose to report this expression of reproducibility as the smallest real difference (SRD) when reporting on evaluative outcome measures (24).

Responsiveness. Furthermore, we propose to quantify the responsiveness of an evaluative outcome measure using an external criterion that classifies the patients as "changed" or "not changed" and a cut-off for the change score on the outcome measure under study. A serious candidate for the cut-off change score is the SRD, but with all possible cut-offs a receiver operating characteristic (ROC) evolves that gives a complete picture of the responsiveness.

Comparison of Change Scores with the SRD. Responsiveness concerns the correct identification of real change according to the external criterion. A perfect evaluative outcome measure will classify all individuals who improve (or deteriorate) according to the external criterion as having changed, while none of the stable patients will be classified as having changed. Individual patients may be labeled as having changed when the change is equal to or larger than the SRD. Consequently, the responsiveness of an evaluative outcome measure can be assessed by taking the proportion of patients with a change score equal to or larger than the SRD for improved (or deteriorated) patients (according to the external criterion). In this context, the measurements for sensitivity to change (for changed patients) and specificity to change (for stable patients) are useful. With the SRD as the cut-off, the specificity to change is, for a normal distribution, equal to 95%.

ROC Curves. Deyo and Centor (9) have drawn an analogy between health status assessment and diagnostic tests. In some ways, determining the responsiveness for evaluative outcome measures is analogous with evaluating the accuracy of a diagnostic test and can be described in terms of its sensitivity and specificity for detecting change as identified by the external criterion. The ROC curve is a graph of the true-positive rate (sensitivity) versus the false-positive rate (1 − specificity) for all possible cut-off change scores (9;10). Using the SRD as cut-off change score is, of course, only one of many possibilities. Under the assumption that sensitivity to change is equally important to specificity to change, the point most upper left in the diagram represents the optimal cut-off change score. Responsiveness can also be quantified using the area under the ROC curve, which combines sensitivity and specificity for all possible cut-off change scores (9;18). The area under the curve (AUC) can be interpreted as the probability of correctly discriminating between improved and nonimproved patients (18). This area theoretically varies from 0.5 (no accuracy in discriminating improved from nonimproved patients) to 1.0 (perfect accuracy). It is important to keep in mind that the AUC is a summary statistic that gives no information about the shape of the curve. To select the most optimal cut-off change score and to interpret the relationship between the true- and false-positive rate inspection of the ROC curve provides the best information.

Illustration

We have used data from a randomized clinical trial on the efficacy of traction for low back pain (3) as an illustration of our proposed methods to assess the responsiveness of evaluative outcome measures. Patients were selected for this trial if they had suffered from nonspecific low back pain for at least 6 weeks.

In this study population, we compared the responsiveness of two well-known instruments for evaluating functional status in low back pain patients: the Oswestry disability

questionnaire (Oswestry) (11) and the Roland disability questionnaire (RDQ) (25). The RDQ was assessed for all 151 patients participating in the trial. We asked the last 81 patients we recruited to also fill out the Oswestry at baseline and 5 weeks later, after the treatment period. The analyses below are limited to this subgroup.

Outcome Measures

The Oswestry disability questionnaire covers 10 activities of daily living that can be hampered by low back pain (11). The Roland disability questionnaire was derived from the Sickness Impact Profile (SIP) by choosing 24 items relevant for low back pain (25).

The external criterion for clinically relevant change was derived from the global perceived effect assessed by the patient on a seven-point Likert scale (1 = completely recovered, 7 = vastly worsened). We asked the patients in what way their low back pain had changed during the last 5 weeks. If a patient indicated complete recovery or much improvement, we classified the patient as having improved, while we classified a slight improvement, no change, or a slight deterioration as being stable. Patients who assessed themselves as much worsened and vastly worsened were classified as having deteriorated. In a secondary analysis we evaluated the influence on responsiveness of a less stringent external criterion. In this analysis the patients who only slightly improved or deteriorated were also defined as having a clinically relevant change.

Data Analyses

The data were analyzed using the Statistical Package for Social Sciences (SPSS 7.0) for Windows 98. The sum score of the Oswestry varies from 0–100, while the sum score of the RDQ varies from 0–24. To enable a better comparison of the sum scores, we presented the standardized RDQ sum score of 0–100 by multiplying each sum score by 100/24. The outcomes of both functional status questionnaires appeared to be normally distributed at baseline. We calculated change scores by subtracting the score after 5 weeks from the baseline score. Thus, a positive change score indicated an improvement, and a negative difference, a deterioration. These change scores appeared to be normally distributed.

RESULTS

Scores on the Instruments

All 81 participants had complete data on the baseline and posttreatment (after 5 weeks) measurements. Through self-assessment of the global perceived effect on a seven-point scale, 6 patients (7%) rated themselves as being completely recovered, 32 patients (40%) as much improved, 19 (24%) as slightly improved, and 15 (19%) reported no change. There were five patients who indicated deterioration (much worsened or vastly worsened). Because of this small number, we evaluated only the improved and stable patients and excluded from further analyses patients whose condition had deteriorated.

Using the stringent external criterion, 38 patients had improved (completely recovered and much improved) and 38 were stable. The 38 stable patients were used to calculate the background noise. Table 1 shows the mean scores with the standard deviations of the improved and stable patients at baseline and posttreatment. The functional status instruments registered different mean scores at baseline. The scores of the Oswestry were lower than the RDQ scores. The baseline scores of the improved and stable group were similar for each instrument.

Reproducibility

Table 2 shows the mean change scores and corresponding standard deviations for the stable patients and the SRD. The mean change score for the stable patients for the Oswestry

Table 1. Mean Sum Scores and Standard Deviations ($SD_{between}$) at Baseline and Post-treatment and the Mean Change Scores in the Improved and Stable Patients for Both Functional Status Instruments

	Baseline	Post-treatment	Change scores
Oswestry (0–100)			
Improved (n = 38)	26.3 (13.5)	14.3 (15.1)	11.9
Stable (n = 38)	29.1 (15.2)	29.5 (17.4)	−0.4
Roland (0–100)			
Improved (n = 38)	50.4 (19.4)	17.9 (17.6)	32.6
Stable (n = 38)	49.3 (21.3)	44.2 (22.8)	5.1

Table 2. Mean Change Scores, Corresponding Standard Deviations, and SRD in Stable Patients

	Change scores	$SD_{change\ ind}$	SRD
Oswestry (0–100)			
Stable (n = 38)	−0.4	9.2	18.0
Roland (0–100)			
Stable (n = 38)	5.1	12.5	24.5

$SRD = 1.96 * SD_{change\ ind}$.

(−0.4) was lower than for the RDQ (5.1). The $SD_{change\ ind}$ was lower for the Oswestry than for the RDQ, and of course the same held for the SRD, which is based on the SD. The SRD was 18.0 for the Oswestry and 24.5 for the RDQ. This means that an individual had to change at least 18.0 points on the Oswestry or at least 24.5 on the RDQ to be judged as having really changed. Consequently, the Oswestry is slightly more reproducible than the RDQ.

Responsiveness

Comparison of Change Scores with the SRD. The change score for each patient was compared with the SRD. Table 3 shows that 26% of the patients who had improved according to the external criterion had a change score higher than the SRD (18.0) on the Oswestry. For the RDQ, 63% of the patients who had improved according to the external criterion had a change score higher than the SRD (24.5). In other words, using the SRD as the cut-off change score, the sensitivity to change of the Oswestry was 26% and of the RDQ, 63%.

For the Oswestry, none of the patients in the stable group had a change score higher than the SRD. But for the RDQ, two of the stable patients (5%) had a change score higher than the SRD. In other words, using the SRD as the cut-off change score, the specificity to change of the Oswestry was 100% and of the RDQ 95%. These values were expected because the specificity to change is by definition around 95% if the SRD is taken as the cut-off change score.

ROC Curves. Figure 2 shows the ROC curves for the two functional status instruments. For both instruments, the curve lies above the diagonal, implying some responsiveness. The curve for the RDQ is clearly closer to the upper left than the curve for the Oswestry. Consequently, the AUC for the RDQ was higher (0.93) than for the Oswestry (0.76). The RDQ turned out to be superior to the Oswestry irrespective of the chosen cut-off change scores.

Table 3. Percentage of Patients Who Scored More Than the SRD Among the Improved and Stable Patients for Both Functional Status Instruments

	% improved patients sensitivity to change	% stable patients 1 − specificity to change
Oswestry >SRD	26%	0%
Roland >SRD	63%	5%

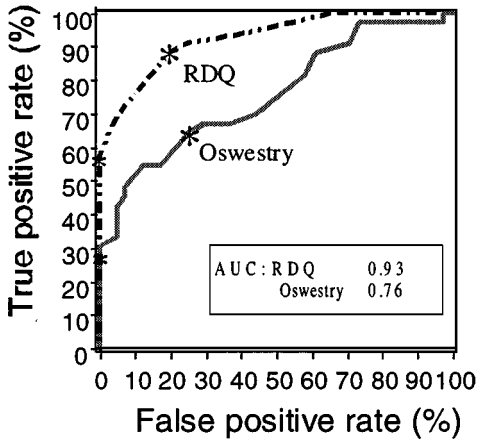

Figure 2. ROC curves of the change scores for the Roland disability questionnaire and Oswestry disability questionnaire (n = 76). The SRD (*) and optimal cut-off points (*) of both questionnaires are plotted in the figure.

The importance of false-positive and false-negative errors may vary depending on the situation. If we assume that sensitivity to change is equally important to specificity to change, the point most upper left in the diagram represents the optimal cut-off change score. Changes that correspond with the best cut-off points in this study were 5 to 6 points of the maximum of 100 points change for the Oswestry and 18 to 20 points of the maximum of 100 points change for the standardized RDQ (Figure 2). With the use of the ROC curve, every user can select the best cut-off change score for their situation.

In the first method, we used the SRDs as the cut-off change score. If a patient changed more than this, the change would be larger than expected due to chance. We plotted the corresponding true- and false-positive rates in the ROC curves, which were 26% and 0%, respectively, for the Oswestry and 63% and 5%, respectively, for the RDQ.

Secondary Analyses with a Less Stringent External Criterion

When using the less stringent external criterion in the secondary analysis, 57 patients improved (completely recovered, much improved, and slightly improved) and 15 were stable. The corresponding SRD was 17.2 points for the Oswestry and 33.3 points for the RDQ. We subsequently compared the change score for each patient with these SRDs. It appeared that 21% of the patients who improved according to the less stringent external criterion had a change score higher than the SRD on the Oswestry compared with 25% on the RDQ. In the stable group one patient (7%) had a change score higher than the SRD on the Oswestry compared with none of the patients on the RDQ. If we compare these results with the results of the stringent external criterion, the sensitivity to change of both

https://doi.org/10.1017/S0266462301107038 Published online by Cambridge University Press

instruments decreased, more for the RDQ than for the Oswestry, while the specificity to change of both instruments remained as expected at about 95%.

Using the less stringent external criterion, the ROC curve for the RDQ appeared to be still closer to the upper left than the curve for the Oswestry. The AUC for both instruments decreased, from 0.93 to 0.82 for the RDQ and from 0.76 to 0.73 for the Oswestry.

DISCUSSION

A large number of different methods to assess reproducibility and responsiveness of evaluative outcome measures are described in the literature (23;26). However, there is as yet no consensus on the most appropriate strategies.

In evaluating responsiveness, it is important to evaluate both the ability of an instrument to detect a real change in health status (sensitivity to change) and the ability to detect absence of change when there is no real change (specificity to change). With the methods we propose it is possible to evaluate both aspects of responsiveness. Our illustration showed that both the Oswestry and RDQ were specific to change but that the RDQ was substantially more sensitive to change than the Oswestry.

Another method used to quantify responsiveness is the calculation of effect size statistics, which relate the magnitude of change (the signal) to the variability in score (background noise) (8;15;17). We used percentages above or below the cut-off instead of the relative magnitude of change. In our opinion this gives a more complete picture and is less sensitive to outliers or extreme changes.

For the method we propose to assess responsiveness, an explicit dichotomous external criterion is necessary to define the minimum change that is considered to be clinically relevant. In our example we used global perceived effect as an external criterion. It is an all-encompassing measure for improvement that includes pain, functional status, and other aspects that patients perceive to be important. Most physicians would be reluctant to label a patient as improved or deteriorated contrary to this personal assessment. From both the patients' and the clinicians' viewpoints, it is relevant and sensible to ask the patient to assess his or her perceived benefit (6;13). This is no perfect gold standard that defines whether a patient has changed. We consider it a surrogate criterion that does not precisely define the smallest amount of change that is clinically relevant. Consequently, the background noise estimation based on this surrogate criterion and the real background noise will typically differ. According to the surrogate external criterion, the no change section of Figure 1 will be defined too wide, which has several consequences. Using a surrogate external criterion, we are not able to assess the absolute responsiveness of evaluative outcome measures, as the size of the statistics for reproducibility and responsiveness will depend strongly on the choice of the external criterion. Despite these complexities, a surrogate external criterion can be used to compare the reproducibility and responsiveness of different evaluative outcome measures in the same study population (2;4;7;8;22;26;29;30;31).

The choice of a cut-off point for the external criterion to determine clinically relevant change is difficult and often arbitrary (23). We performed two analyses. In the first stringent analysis, we classified complete recovery or much improvement as improved. In a secondary analysis, the patients who only slightly improved were also defined as having a clinically relevant improvement. In this analysis we saw that the sensitivity to change of both instruments decreased, although the ranking of the outcome measures stayed the same. We preferred the use of the stringent external criterion because we wanted to measure a clinically relevant difference and we wanted to anticipate socially desired answers. Our example illustrates that when evaluating the relative responsiveness, it is important to compare the instruments against the same external criterion. It is also informative to compare the same instruments against several external criteria in the same study population (7). If

results are consistent on the basis of several external criteria, confidence increases about the correct ranking of the responsiveness of the outcome measures (10).

Analogues to diagnostic studies, the responsiveness of instruments strongly depends on the characteristics of the study population in which the evaluation takes place. There is, for example, some evidence indicating that the Oswestry is more responsive for patients with severe low back pain, while the RDQ is more responsive for patients with less severe low back pain (1).

In clinical trials, the aim is to compare change over time between groups of patients. The methods proposed in this article can also be applied to groups of patients. To quantify reproducibility of an evaluative outcome measure in a group of patients, the $SD_{change\ ind}$ has to be divided by $\sqrt{n}$. To evaluate whether the observed change in a group of patients is larger than expected due to background noise alone, the relevant statistics ($SD_{change\ ind}$) / $\sqrt{n}$) can be multiplied by 1.96. Observed change of at least this size implies a real change with 95% confidence. To evaluate the responsiveness at group level, the mean change score in the improved and stable group has to be compared with this SRD_{group}. In our example with a group size of 38, the SRD_{group} of the Oswestry was 2.9 and of the RDQ, 4.0. The fact that the SRD_{group} is much smaller than the corresponding $SRD_{individual}$ implies that the same evaluative outcome measure is much better at detecting change at group level.

The proposed statistics for quantifying reproducibility and responsiveness for individuals are independent of the number of patients in the study. Only the precision of their estimation is dependent on the size of the group at issue. On the contrary, for groups of patients, the size of the group has a direct influence on the magnitude of the SRD_{group}. This has to be taken into account when comparing statistics of different study populations.

In conclusion, in this article we presented the use of SRDs and ROC curves for quantifying responsiveness. We started with basic notions and arrived at, in our opinion, methods that are both understandable and useful.

REFERENCES

1. Baker JD, Pynsent PB, Fairbank JCT. The Oswestry disability index revisited: Its reliability, repeatability and validity, and a comparison with the St. Thomas's disability index. In: Roland MO, Jenner JR, eds. *Back pain: New approaches to rehabilitation and education*. Manchester: University Press; 1989:174-186.
2. Beaton DE, Hogg-Johnson S, Bombardier C. Evaluating changes in health status: Reliability and responsiveness of five generic health status measures in workers with musculoskeletal disorders. *J Clin Epidemiol*. 1997;50:79-93.
3. Beurskens AJHM, de Vet HCW, Köke AJA, et al. Efficacy of traction for non-specific low back pain: A randomized clinical trial. *Lancet*. 1995;346:1596-1600.
4. Beurskens AJHM, de Vet HCW, Köke AJA. Responsiveness of functional status in low back pain: A comparison of different instruments. *Pain*. 1996;65:71-76.
5. Bombardier C, Tugwell P. Methodological considerations in functional assessment. *J Rheumatol*. 1987;14(suppl 15):6-10.
6. Bombardier C, Tugwell P, Sinclair A, et al. Preference for endpoint measures in clinical trials: Results of structured workshops. *J Rheumatol*. 1982;9:798-801.
7. Bronfort G, Bouter LM. Responsiveness of general health status in chronic pain: A comparison of the COOP Charts and the SF-36. *Pain*. 1999;83:201-209.
8. Cohen J. *Statistical power analysis for the behavioral sciences*. 2nd ed. New York: Academic Press; 1988:1-27.
9. Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: An analogy to diagnostic test performance. *J Chron Dis*. 1986;39:897-906.
10. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures: Statistics and strategies for evaluation. *Control Clin Trials*. 1991;12:142S-158S.
11. Fairbank JCT, Couper J, Davies JB, O'Brien JP. The Oswestry low back pain disability questionnaire. *Physiotherapy*. 1980;66:271-273.

12. Feinstein AR. *Clinimetrics.* New Haven: Yale University Press; 1987.
13. Fries JF. Toward an understanding of patient outcome measurement. *Arthritis Rheum.* 1983;26:697-704.
14. Goldsmith CH, Boers M, Bombardier C, Tugwell P. Criteria for clinically important changes in outcomes: Development, scoring and evaluation of rheumatoid arthritis patient and trial profiles. *J Rheumatol.* 1993;20:561-565.
15. Guyatt G, Walter S, Norman G. Measuring change over time: Assessing the usefulness of evaluative instruments. *J Chron Dis.* 1987;40:171-178.
16. Guyatt GH, Juniper EF, Walter SD, Griffith LE, Goldstein RS. Interpreting treatment effects in randomised trials. *BMJ.* 1998;316:690-693.
17. Guyatt GH, Kirschner B, Jaeschke R. Measuring health status: What are the necessary measurement properties? *J Clin Epidemiol*. 1992;45:1341-1345.
18. Hanley JA, McNeil BJ. The meaning and use of the area under the receiver operating characteristic (ROC) curve. *Radiology.* 1982;143:29-36.
19. Hays RD, Hadorn D. Responsiveness to change: An aspect of validity, not a separate dimension. *Qual Life Res.* 1992;1:73-75.
20. Juniper EF, Guyatt GH, Willan A, Griffith LE. Determining a minimal important change in a disease-specific quality of life questionnaire. *J Clin Epidemiol.* 1994;47:81-87.
21. Kirshner B, Guyatt GH. A methodological framework for assessing health indices. *J Chron Dis.* 1985;38:27-36.
22. Liang MH. Evaluating measurement responsiveness. *J Rheumatol.* 1995;22:1191-1192.
23. Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: The lesson of Cronbach. *J Clin Epidemiol.* 1997;50:869-879.
24. Roebroeck ME, Harlaar J, Lankhorst GJ. The application of generalizability theory to reliability assessment: An illustration using isometric force measurements. *Phys Ther.* 1993;73:386-401.
25. 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.* 1983;8:141-144.
26. Stratford PW, Binkley JM, Riddle DL. Health status measures: Strategies and analytic methods for assessing change scores. *Phys Ther.* 1996;76:1109-1123.
27. Streiner DL, Norman GR. *Health measurement scales: A practical guide to their development and use.* Oxford: University Press; 1995.
28. Stucki G, Liang MH, Fossel AH, Katz JN. Relative responsiveness of condition-specific and generic health status measures in degenerative lumbar spinal stenosis. *J Clin Epidemiol.* 1995;48:1369-1378.
29. van Bennekom CAM, Jelles F, Lankhorst GJ, Bouter LM. Responsiveness of the Rehabilitation Activities Profile and the Barthel Index. *J Clin Epidemiol*. 1996;49:39-44.
30. van der Heijden GJMG, Leffers P, Bouter LM. Shoulder Disability Questionnaire: Construction and responsiveness of a functional status measure. *J Clin Epidemiol.* 2000;53:29-38.
31. van der Windt DAWM, van der Heijden GJMG, de Winter AF. The responsiveness of the shoulder disability questionnaire. *Ann Rheum Dis.* 1998;57:82-87.
32. Wells GA, Tugwell P, Kraag GR, et al. Minimum important difference between patients with rheumatoid arthritis: The patient's perspective. *J Rheumatol.* 1993;20:557-560.
33. Wright JG, Young NL. A comparison of different indices of responsiveness. *J Clin Epidemiol.* 1997;50:239-246.