Responsiveness of Outcome Measures

María Jesús García de Yébenes Prous,[a] Francisco Rodríguez Salvanés,[b] and Loreto Carmona Ortells[a]

[a]Unidad de Investigación, Fundación Española de Reumatología, Madrid, Spain
[b]Agencia Laín Entralgo, Madrid, Spain

In medical research, particularly in the field of rheumatology, there is great interest about the concept of responsiveness of outcome measures as a sign of changes in the patient's health status. However, the terminology surrounding this concept and the methods of analysis are confusing and lacking in consensus. We present a review about the concept and analysis of responsiveness taking into account both, the characteristics of the responsiveness and the type of design and predictable change in the sample being studied.

Key words: Responsiveness. Outcome measures. Responsiveness statistics.

Sensibilidad al cambio de las medidas de desenlace

En la investigación médica, y en especial en el ámbito de la reumatología, hay gran interés sobre el concepto de sensibilidad al cambio de un instrumento de medida como posible reflejo de modificaciones en la situación clínica del paciente. Sin embargo, la terminología de este concepto y su metodología de análisis están rodeadas de confusión y falta de consenso.
Se presenta un trabajo de revisión sobre el concepto y el análisis de la sensibilidad al cambio teniendo en cuenta tanto las características de la sensibilidad como el tipo de diseño y cambio previsible en la muestra en estudio.

Palabras clave: Sensibilidad al cambio. Medidas de desenlace. Estadísticos de sensibilidad al cambio.

Correspondence: Dra. M.J. García de Yébenes y Prous.
Unidad de Investigación. Fundación Española de Reumatología.
Marqués del Duero, 5, 1.º A. 28001 Madrid. España.
E-mail: mjesus.garciadeyebenes@ser.es

Manuscript received June 27, 2008; accepted for publication July 24, 2008.

Responsiveness to Change

In general, clinicians and researches evaluate the quality of functional measurement scales in relation to their effectiveness and validity. However, responsiveness to change can also be considered a fundamental characteristic of the evaluation instruments, designed to measure a longitudinal change through time.[1] There is no consensus in the literature on the concept of responsiveness of an instrument or the form in which it should be measured. In a review of medical literature, Terwee et al[2] encountered different definitions and 31 different ways to measure responsiveness to change. This lack of consensus has originated a great proliferation of statistics, and it is not uncommon for some researchers to employ several simultaneously in the same article,[3,4] making it difficult or even preventing the comparison between measures in these studies.[5] Uncertainty on the type of designs and the methods of analysis can be due, in great measure, to the absence of a responsiveness parameter regarding the health status of the patients.[6]
Responsiveness to change is the degree with which different results are obtained after repeated applications of the same instrument when a real change in the health status has occured.[7] Therefore, it can be defined as the capacity of an instrument to detect change. This characteristic has generated a great deal of interest, because change in a measure can reflect a modification of the patients clinical situation, something important for intervention studies. The study of this dimension requires a standard that indicates clinical change. The traditional method used to evaluate the characteristics of change is the retrospective and general evaluation of the change produced.[7,8] This method consists in not only performing the test under evaluation in the follow-up visits, but asking the patient for his or her impression on the change. This unique item of general retrospective analysis is used later to evaluate the capacity of the instrument to detect change.[9] However, this procedure has several inconveniences: *a)* the metric properties (internal coherence and validity) of a single item of general retrospective evaluation are inferior to those of the instrument or multiple item questionnaire under study; *b)* from a psychological standpoint, the general retrospective evaluation is difficult and subjective; and *c)* the use of the general retrospective evaluation is based on the presumption that it is a measure independent from

the study instrument, leading to the fact that measurement errors are not correlated. However, any exacerbation or remission at the moment of observation can influence the patient evaluation. Correlation of the measurement error between the general retrospective evaluation and the test overestimated the real relationship between them. Therefore, the evaluation of responsiveness to change based on a general retrospective evaluation has little value for studying the capacity of an instrument to detect change.[4,8]

An alternative to general retrospective evaluation is the prognostic evaluation of change, which consists in an a priori statement of the characteristics of the change that is to be produced in the sample. This procedure is not the subject of recall or correlation errors, but may depend on the capacity of the evaluator to estimate the exact extension of the change that can be produced. It has been used, for example, in randomized trials in which known efficacy interventions are compared to placebo, or in cohort studies in which a known prognostic variable is used to classify patients into groups that will probably change in a different manner.[8]

Responsiveness to change depends on the population of patients under study and the scenario in which the measurement instrument is applied. For example, the variation in scores for change will be larger in a heterogeneous population compared to a homogeneous one; in the same way, when the intervention is very effective in some of the patients but not in others, it will be more important than when it has little efficacy in all of the patients, even though responsiveness to change is the same in both cases.[2] Therefore, the choice of an analytic method and the respective responsiveness to change coefficients fundamentally depend on the characteristics of the sample, especially on the type of design and the change expected[8,10]:

– A single group: homogeneous change. The sample is formed by a single group in which patients presumably will change more or less in the same manner between 2 moments. Coefficients used are based in the homogeneous change between patients

– A single group: heterogeneous change. The sample is also formed by a single group but, in contrast to the former one, it is foreseeable that the patients change differently from one another. Coefficients are based on a correlation analysis

– Different subgroups: mixed change. In this case the sample is constituted by 2 or more subgroups of patients that change in a different way between 2 moments. It shares characteristics with the other 2: change will be different in the subgroups of patients defined in the sample, making it a type 2 design, but also, in each group, change will probably be homogeneous, conferring it with characteristics of the first design. coefficients are based on differences in change between subgroups

Some authors use other classification systems for responsiveness to change. In this context, Husted et al distinguish 2 great types of responsiveness to change: internal and external.[5,11]

Sensitivity to internal change is defined as the capacity of a measure to change in a determined lapse of time; it refers to the possibility of detecting any kind of statistical change. For example, a single group of patients is studied before and after the application of an effective treatment. Sensitivity to change will depend both on the treatment used as on the measure of the response used to determine therapeutic efficacy.

External responsiveness to change is defined as the degree with which changes in a measurement with time are related to the corresponding changes in a standard measurement of the health status. This dimension of responsiveness to change is associated to the concept of clinical relevance and consists in the property of a measurement to capture a clinically important change. In contrast to the internal, the fundamental difference is not in the measure itself but in the relationship between change in the measurement and change in the external standard. If this relationship is important, the measurement adequately captures changes in the external standard. It is accepted that changes in the standard are an indication of a modification of the patient's situation. Therefore, that change in the measurement is able to capture change in the standard could indicate a modification in the subjects clinical situation. Responsiveness to external change will only depend on the choice of external standard and not on the treatment under study. Therefore, this type of responsiveness to change will be a property of the measurement instrument.

Table 1 presents a classification of the responsiveness to change in relation to its characteristics and the type of design/change likely.

We now will present the most common statistical tests for the evaluation of responsiveness to change, both in relation to the study design and the characteristics of likely change as for the classification of internal and external responsiveness to change.

Homogeneous Change

As has been noted, this design and its corresponding analysis are based on the premise that the sample is formed by a single group of patients who will most likely change in the same way during the study period. What is important are not that the factors upon which change depends (natural history or the application of an effective treatment), but that the magnitude of change is homogeneous among patients.

This design would measure the previously mentioned "responsiveness to internal change," because in truth, what is evaluated is the capacity of the measure to change during

TABLE 1. Responsiveness to Change According to its Characteristics and the Type of Design

Design/ Type of Change	Characteristics	
	Internal	External
Homogeneous	*t* test related data Size of the standardized test Mean standardized response (MSR) Guyatt statistic	
Contrast between subgroups		ROC curves ANOVA repeated measurements Norman *S*
Contrast between patients		Correlation Regression models

a concrete lapse in a sample of patients who have shown improvement, for example, after the application of a treatment ok known effectiveness.[5]
The most commonly used statistical tests are based on the mean or general change in group, and among them we should mention the following.

t Test for Related Data

The *t* test proves the absence of change hypothesis of the mean response of a measure between 2 moments. Because it is a design with repeated measurements in the same subject, *t* is used for related data.

$$t = \text{mean difference} / (SD_{difference} / \sqrt{n})$$

The *t* test is centered on the statistical significance of observed change which depends, evidently, on the magnitude of change, but also on the size of the sample and the variability of the measure. This makes it a weak statistical test for evaluating responsiveness to change. From a methodological standpoint, it is more useful to employ coefficients in which both the magnitude to change as well as its variability intervene.

Statistical Tests Based on the Size of the Effect

In contrast with *t* tests, statistics based on the size of the effect provide direct information on the magnitude of change, expressed as a variation. Therefore, statistics related to size of effect measure the relationship between the magnitude of change (signal) and variability (background noise).

Size of the Standardized Effect

Statistics from this group, frequently employed, are the size of the standardized effect or the relationship between the mean of the differences in the baseline scores and the follow-up, divided by the standard deviation (SD) of the baseline score.[3]

$$\text{Size of the effect} = \text{mean change} / SD_{\text{baseline measurement}}$$

A size of the effect of 0.20 indicates that the change is approximately one-fifth of the SD of the baseline measurement and is defined as small. Values of 0.50 are considered as moderate while those 0.80 and larger are important. A limitation of the size of the effect is that we are unable to know if it reflects a real change or just the variability of the baseline score.

Mean Standardized Response (MSR)

It is another test based on the size of the effect. It is calculated by dividing the mean change by its SD, reflecting the variability of the change scores. Therefore, if change has an elevated variability with respect to its mean, a low MSR will be obtained.

$$MSR = \text{mean change} / SD_{change}$$

The mean standardized response is independent from the size of the sample and, in addition takes into account the variability to change, making it the most adequate statistic for the study of responsiveness in this type of design. As happens with the size of the effect, 0.20, 0.50, and 0.80 values indicate instruments with poor, moderate, and elevated responsiveness to change.[5,8]
The value of MSR must be accompanied by a confidence interval. Some authors calculate this interval assuming that the differences in score follow a normal mean distribution of 0 and $DE=1/\sqrt{n}$, although others use resampling methods such as Jacknife.[12]

Guyatt Statistics

Guyatt statistic, also known as *responsiveness statistic,* is also based on the size of the effect and consists in the coefficient between the minimum clinically important difference (MCID) and the error of the quadratic mean in a variance analysis of repeated measures in clinically stable patients (MSE).

$$\text{Guyatt} = MCID/\sqrt{(2\times MSE)}$$

In the case that only 2 observations are available (before and after an intervention), the denominator would be the

SD of the individual change scores in stable patients. MCID is defined as the smallest difference between the baseline and follow-up scores that are associated to an important clinical effect in a group of patients. In general, MCID is established by relating the changes in the measurements produced by a clinical change standard, such as the patients own evaluation. MCID reflects the magnitude in change in the measurement associated with an arbitrary definition of the minimally important change in the external standard. There are multiple definitions of MCID; one of the most commonly employed is the difference in mean change between patients that show some improvement and those who do not show any changes in their health status. The denominator adjusts possible sudden changes due to a measurement error or learning effects.

Guyatt statistics is a form of detecting modifications in the different outcome measures of the random changes presented by the patients who do not show detectable improvements.[6] In this manner, a measurement will be responsive to change if it is capable of detecting MCID larger than any sudden change.[5,13] This test is seldom used, at least in rheumatology, due to the difficulties inherent to the definition of minimally clinical important change.[13]

Responsiveness to internal change, or the capacity to detect a statistical change in measurement, can be affected by different structural parameters such as the type of scale, the scoring system and the number of items related with the "signal" and the "noise." An elevated number of items tends to increase the responsiveness if they are not redundant. The continuous scales that cover the complete spectrum of an outcome, from the mild to the severe forms, generally avoid the "flooring" and "ceiling" effects and increase responsiveness. Scales with gradual scoring systems also have a larger responsiveness to change that dichotomous ones. Lastly, trustworthiness is a factor to take into account because it intervenes in the magnitude of the denominator of the statistical tests based on the size of the effect.[11]

Excepting Guyatt, none of these tests relates the change of the scale under study with that produced in a measurement of the clinical situation. Any change observed is attributed to a modification in the status of the patient but, in truth, statistically significant changes can be observed in the measurement without having a relevant modification in the clinical status of the patient.[4,5] Therefore, they are only useful to evaluate the intrinsic or internal capacity of the instrument for change. On the other hand, although Guyatt's test is related to a standard in clinical change, when the sample of the patients is stable, it is not allows for the differentiation between different amounts of change (worsening, improvement). These limitations make the homogeneous design (responsiveness to internal change) the weakest one, because it does not allow to discriminate between different magnitudes of change.[8]

Mixed Design: Contrast in Subgroups

This design shares characteristics from the other 2. The sample is formed by subgroups of patients that change in a different manner (heterogeneous change), but in each subgroup the change in the patients is uniform (homogeneous change). It would be equivalent to responsiveness to external change.

This type of design can use different change coefficients.

ROC Curves

The use of receiver operating characteristics (ROC) curves as an evaluation method of responsiveness to change in rheumatology was initially proposed by Deyo et al[14] in 1986.

An outcome measure can be described, in a similar way to a diagnostic test, by its clinical capacity to correctly identify individuals who present an important clinical change. In order to perform this analysis, it is necessary to have an external standard for change. In this way, responsiveness to change is described in terms of sensitivity (the probability that a measure will correctly classify patients who had changes in an external indicator of change) and specificity (probability that the measure will correctly classify patients who did not show changes in the external standard). In other words, it attemps to evaluate the capacity of the measurement to reflect differences in change between groups with respect to the external standard (improvement/absence of improvement).

The area under the curve (AUC) expresses the discriminative capacity of the instrument or the probability of correctly classifying both the patients who show improvement (or worsening) and the ones that do not improve (or do not worsen). ROC curves provide a general vision of the relationship between a measure and an external standard for change. The main disadvantages are that the classification variable (standard) must be dichotomized, making it possible to loose information on the magnitude of change, and that it requires independent analysis to define responsiveness to change for improvement and worsening.[8,15]

Differences in the Measures of Change Scores

The study of the differences in the means of the scores between the different subgroups of the sample can be performed through an analysis of variance (ANOVA) of repeated measures with an intrapatient factor (a measurement opportunity with 2 levels) and a factor between patients (magnitude of change with 2 or more levels: improvement/no improvement). The results of ANOVA allow for obtaining the extension with which the subgroups differ regarding change, represented by the

significance of the end of interaction (group X opportunity of measurement).

Norman's *S*

Norman's *S* is a test derived from an analysis of variance of repeated measurements and represents the relationship between variance of interaction and the sum of this and variance of error.[16]

Heterogeneous Composition: Contrast Between Patients

In this design, as in the homogeneous, the sample is formed by a single group. However, patients are not homogeneous among them regarding change, and it is foreseeable that they change in a different manner. An essential aspect is that an external standard is applied with change scores that are compared with those of the measurement in question. Therefore, responsiveness to change will depend on the choice of a standard and not the interventions performed. In these cases, responsiveness to external change is what is being evaluated. The correlation coefficient and regression models are the most common tests used in these designs.

Correlation

Pearson's correlation coefficient is calculated by the change scores between 2 measures (*x and y*).

$$r_{xy}=\sum_{i=1}^{n}(D_{xi}-\overline{D}_x)(D_{yi}-\overline{D}_x)/n$$

The correlation coefficient indicates how the 2 measurements change simultaneously and their values oscillate between −1 and +1. In general, x is the measure of the study and y is the specific clinical result (for example, a functional capacity score). If r_{xy} is close to 1, the measurement captures the information contained in y, in other words, the instrument responds to changes in the clinical result.

The main limitations of the coefficient of correlation derive from its variation in the response to the selection of values (for example, elimination of outliers) and the presumption of linearity between both variables, when in truth this relationship can be non-linear.[8]

Regression Models

An interesting aspect of this design is to examine if the changes in a measurement are capable of predicting changes in the other. This analysis can be performed through regression models.

$$D_{yi}=\alpha+\beta d_{xi}+e_i$$

The parameter represents mean change in the clinical result (D_y) in the absence of change in the evaluated instrument (d_x=0) for each unit of change in the instrument (d_x). β values close to 0 indicate that important changes in *x* may not be accompanied by changes in y, while elevated values for β indicate that the changes associated in y will also be important. This model can be generalized by adding more variables that predict change, even the baseline clinical situation (Y_1). In the same manner, it is possible to standardize the coefficients for the purposes of comparison with other studies.[5]

The main limitations of this design are related to the choice of an external standard. In spite of an elevated correlation coefficient, the standard may not capture all of the observed changes in patients and on the other hand, the standard and the instrument could be measuring different aspects of one same concept.

In order to facilitate comprehension of all of these indexes, we will present an example. Imagine that a study is performed in which a health scale, with a scoring range of 0 to 25 is applied, to a group of 20 patients in a baseline visit and a during the baseline visit and after 6 months of follow-up. During the follow-up visit, the patient is asked for a general evaluation of change experimented on a scale of −5 to 5. According to this general evaluation, a new variable is created, denominated as group, which classifies patients according to wether they experienced change or not (values 1 and 0, respectively) using a cutpoint of = 4 (Table 2).

With this data, and making no presumptions on the expected change, different types of analysis and statistics are presented according to the size of the responsiveness to change (internal and external) and the different characteristics of foreseeable change (homogeneous, heterogeneous between groups, heterogeneous between patients).

If we assume that the sample is formed by a single group of patients who will probably change more or less in the same way, the test used to calculate will be and MSR=0.69(3.4/4.9); in other words, the measurement in question will have a moderate responsiveness to change. This test does not inform us of the possible differences in change between the 2 defined categories defined by the group variable.

Let us suppose, on the contrary, that the sample is formed by 2 subgroups of patients in which change is produced in a different manner. In the previous example, these 2 groups would be represented by the values 0 and 1 from the group variable. In this case, responsiveness to change could be studied through the construction of ROC curves between differences in the scoring and the change category

(group). The results offer a AUC of 0.869 (0.715-1), with these sensitivity and specificity values of 78% and 73%, respectively, for a difference between the baseline score and the follow-up score = 5 units (Table 3). The main limitation of this procedure could be related to the need to dichotomize the results of the external standard. In this form, only the "improvement/no improvement" categories have been included, although some patients may have worsened and those would have required a new classification.

Another way of approaching the contrast between the subgroups of the sample is to perform an ANOVA of the repeated measures with an intrapatient factor (opportunity of measurement) and another between patients (type of change). The term of interaction opportunity of measurement-type of change informs us of the extension with which the 2 groups change in a different manner. In our example, the term interaction is significant, making the magnitude of change different between both groups (Table 4).

Imagine, finally, that our sample is formed by a heterogeneous group of individuals who will foreseeably change in a different way. In this situation a correlation analysis between the change produced in the measure in question and an external standard can be used. In our example, the Spearman correlation coefficient between the change produced (difference between the baseline and the follow-up score) and the external standard (general evaluation of the patient) is 0.71.

In responsiveness to change studies it is important to take into account 2 dimensions, internal and external, and the use of an adequate design that allows for the evaluation of possible differences in change between groups or patients.[2,8,17]

In this sense, Veehof et al have recently published a study on the psychometric properties of 2 activity indexes in rheumatoid arthritis (RADAI and its abbreviated form) in a cohort of patients with this disease and who started treatment with tumor necrosis factor (TNF) alpha inhibitors. The authors postulated their responsiveness to change study for the reduction in the activity (improvement) subdividing it in its 2 dimensions, internal and external. In the case of internal, MSR is calculates with its confidence interval. For the external, the EULAR criteria for classifying responders and non responders are employed as an external standard.

Differences in change between both groups are not captured by MSR, because they are contained by their variability. Therefore, some authors use ROC curves between change scores and the external criteria used to evaluate the discriminative capacity between both groups. In addition, MSR is calculated within each one of them.

TABLE 2. Data From a Hypothetical Study

Patient	Baseline	Follow-up	Difference	General Evaluation	Group
1	25	18	7	5	1
2	20	20	0	−3	0
3	15	6	9	4	1
4	9	5	4	5	1
5	24	12	12	4	1
6	15	18	−3	−4	0
7	8	6	2	3	0
8	12	6	6	4	1
9	15	10	5	4	1
10	14	7	7	5	1
11	12	12	0	-3	0
12	8	5	3	5	0
13	12	8	4	1	0
14	20	15	5	2	1
15	10	19	−9	−4	0
16	12	5	7	4	1
17	17	10	7	3	1
18	9	3	6	3	0
19	21	21	0	−4	0
20	18	22	−4	0	0
Mean	14.8	11.4	3.4	1.7	
Standard deviation	5.2	6.3	4.9	3.4	

TABLE 3. Diagnostic Performance Curve

Cutpoint	Sensitivity, %	Specificity, %	Correct Classification, %	Reason of Truth +
≥−9	100	0	45	1.0
≥−4	100	9.1	50	1.1
≥−3	100	18.2	55	1.2
≥0	100	27.3	60	1.4
≥2	100	54.5	75	2.2
≥3	100	63.6	80	2.7
≥4	88.9	63.6	75	2.8
≥5	77.8	72.7	75	2.8
≥6	66.7	81.8	75	3.7
≥7	55.6	90.9	75	6.1

TABLE 4. ANOVA of Repeated Measures

Origin	Sum of Squares	gl	Quadratic Measurement	F (*P*)
Between subjects	2.5	1	2.5	
Change	2.5	1	2.5	0.04 (.837)
In the same subjects	238.6	2	119.05	
Visit	115.6	1	115.6	19.11 (.0004)
Visit × change	122.5	1	122.5	20.25 (.0003)
Residual	108.9	18	6.0	

The results show a moderate internal and external responsiveness to change (MSR, 0.76 and 0.80; AUC, 0.77 and 0.78). Responders had important improvements in disease activity (MSR>0.80), while non responders did not show any improvement (MSR<0.20).[18]

Limitations Related With the Use of Inadequate Coefficients

The different measurement methods of responsiveness to change have different objectives, leading to different concluisons.[2] Employing an inadequate coefficient for the type of the design under study can make the signal (real change) of some of the coefficients be included into the noise (variability) of others, passing inadvertently. For example, is responsiveness to change with a MSR on a sample formed by subgroups or patients that will change in a different way, the scores related to change or between patients will be contained in the variability (noise) portion of the MSR, reducing its magnitude. However, in spite of the fact that the signal of change between groups or patients could be contained in the noise of the MSR, it is possible to obtain an MSR different from 0 for different reasons.[8]

In the first place, apart from the punctual estimate, a confidence interval must be calculated for the coefficient used with the objective of evaluating the probability that the estimate will really be different from 0.[19]

In second place, researchers interested in evaluating the capacity of a measurement to detect change select patients who are generally expected to improve. Therefore, the mean change of the sample will be more than 0 even when some patients remain stable or even worsen. When the mean change is more than 0, MSR will be more than 0 even if subgroups or individual patients exist who change in a different way.

In third place, it could occur that in a really homogeneous sample with respect to change, apparent differences in change between patients is observed, represented by an elevated correlation with another measurement. In general, these cases are owed to a presumption on the existence of a correlation between the scores of change in the measurement and the general retrospective evaluation of the patient itself. However, using a retrospective evaluation overestimates the existing correlation between both measures. To understand the mechanism of this apparent association, it is necessary to consider the relationships and presumptions on the observed scores (provided by the patients), real scores (unknown values that represent scores which could be obtained in the absence of measurement errors) and errors in measurement (differences between real and observed scores). In theory, change scores and retrospective evaluation are measuring the same attribute, but it is considered that both measurements are independent and, therefore, measurement errors are not correlated. However, it is unlikely that the errors are independent when it is the patients itself who provides the change scores both in the measurement as in the general evaluation. The consequence is that the observed correlation will be larger than 0 even when the correlation between the real scores is 0.[4] In this sense, Fransen et al[20] compared the responsiveness to change on measurements based on the patient's perception and objective measurements in order to detect flare-ups of rheumatoid arthritis. The results proved a reduced responsiveness to change of the items which were subjective in nature. Therefore, although the responsiveness to change coefficients may be similar, the subjective and objective outcome measures are not interchangeable.

Therefore an as a conclusion, before conducting a responsiveness to change study it is important to know the validity and reproducibility of the instrument in question. In addition, it is fundamental to perform a good design of the sample to analyze, define the characteristics of foreseeable change and elect an external standard adequate in necessary cases. It is important to take into account that the application of inadequate responsiveness to change measurements can produce untrustworthy results.

References

1. Guyatt GH, Kirshner B, Jaeschke R. Measuring health status: what are the necessary measurement properties? J Clin Epidemiol. 1992;45:1341-5.
2. Terwee CB, Dekker FW, Wiersinga WM, Prummel MF, Bossuyt PM. On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation. Qual Life Res. 2003;12:349-62.
3. 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.
4. Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. J Clin Epidemiol. 1997;50:869-79.
5. Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing responsiveness: a critical review and recommendations. J Clin Epidemiol. 2000;53:459-68.
6. Walsh TL, Hanscom B, Lurie JD, Weinstein JN. Is a condition-specific instrument for patients with low back pain/leg symptoms really necessary? The responsiveness of the Oswestry Disability Index, MODEMS, and the SF-36. Spine. 2003;28:607-15.
7. de Vet HC, Bouter LM, Bezemer PD, Beurskens AJ. Reproducibility and responsiveness of evaluative outcome measures. Theoretical considerations illustrated by an empirical example. Int J Technol Assess Health Care. 2001;17:479-87.

8. Stratford PW, Riddle DL. Assessing sensitivity to change: choosing the appropriate change coefficient. Health Qual Life Outcomes. 2005;3:23.
9. Koh ET, Leong KP, Tsou IY, et al. The reliability, validity and sensitivity to change of the Chinese version of SF-36 in oriental patients with rheumatoid arthritis. Rheumatology (Oxford). 2006;45:1023-8.
10. Beaton DE, Bombardier C, Katz JN, Wright JG. A taxonomy for responsiveness. J Clin Epidemiol. 2001;54:1204-17.
11. Corzillius M, Fortin P, Stucki G. Responsiveness and sensitivity to change of SLE disease activity measures. Lupus. 1999;8:655-9.
12. Liang MH, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopedic evaluation. Med Care. 1990;28:632-42.
13. Beaton DE, Boers M, Wells GA. Many faces of the minimal clinically important difference (MCID): a literature review and directions for future research. Curr Opin Rheumatol. 2002;14:109-14.
14. Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. J Chronic Dis. 1986;39:897-906.
15. Stratford PW, Binkley FM, Riddle DL. Health status measures: strategies and analytic methods for assessing change scores. Phys Ther. 1996;76:1109-23.
16. Norman GR. Issues in the use of change scores in randomized trials. J Clin Epidemiol. 1989;42:1097-105.
17. Salaffi F, Stancati A, Neri R, Grassi W, Bombardieri S. Measuring functional disability in early rheumatoid arthritis: the validity, reliability and responsiveness of the Recent-Onset Arthritis Disability (ROAD) index. Clin Exp Rheumatol. 2005;23 Suppl 39:S31-42.
18. Veehof MM, Ten Klooster PM, Taal E, van Riel PL, van de Laar MA. Psychometric properties of the Rheumatoid Arthritis Disease Activity Index (RADAI) in a cohort of consecutive Dutch patients with RA starting anti-tumour necrosis factor treatment. Ann Rheum Dis. 2008;67:789-93.
19. Wright JG, Young NL. A comparison of different indices of responsiveness. J Clin Epidemiol. 1997;50:239-46.
20. Fransen J, Hauselmann H, Michel BA, Caravatti M, Stucki G. Responsiveness of the self-assessed rheumatoid arthritis disease activity index to a flare of disease activity. Arthritis Rheum. 2001;44:53-60.