Global Rating of Change Scales: A Review of Strengths and Weaknesses and Considerations for Design

STEVEN J KAMPER, PT, BAppSc¹; CHRISTOPHER G MAHER, PT, PhD²; GRANT MACKAY, PT, BAppSc³

The question of whether a patient has improved or deteriorated is fundamental to clinical practice¹ and the information gained is used in making decisions regarding prognosis, treatment, and ongoing management. Measurement of outcome is a critical step in the conduct of clinical practice. The expectation that clinicians operate within the bounds of an evidence-based practice model has served to highlight the need for clinicians to understand and use appropriate measurement instruments. The *American Physical Therapy Association's* vision statement² acknowledges this fact, and the need for guidance in the area has been recognized³. In practice, clinicians routinely question their patients as to whether they are better, worse, or the

same⁴; but it is unlikely that many attempt to quantify the magnitude of this change or consider the reliability of the information so gained.

Global rating of change (GRC) scales are very commonly used in clinical research, particularly in the musculoskeletal area⁵⁻¹¹. These scales are designed to quantify a patient's improvement or deterioration over time, usually either to determine the effect of an intervention or to chart the clinical course of a condition. GRC scales ask that a person assess his or her current health status, recall that status at a previous time-point, and then calculate the difference between the two¹². The magnitude of this difference is then scored on a numerical or visual analogue scale. Such scales have been recom-

mended for use as a core outcome measure for chronic pain trials¹³ and have also been advocated for improving the applicability of information from clinical trials to clinical practice¹⁴.

However, reliable and accurate function of a GRC scale places considerable cognitive demand on the patient, and a prominent criticism of the measure is founded in the contention that people are unable to accurately recall prior health states^{15,16}. Clearly, an ability to recall and quantify status at a previous time-point is necessary for proper function of the measure as described above. If the reliability of recall is poor, it may be that the "change" score measured by GRC scales is unduly influenced by the status of the patient at the time of scale administration^{17,18}.

The "global" aspect of the scales is important and distinguishes them from other outcome measures that are typically directed towards one specific dimension of the patient's health status such as pain, disability, work ability, or quality of life. Instead, these scales allow the patients themselves to decide what they consider important, an approach that means that the specific constructs each patient takes into account are unknown and may vary. The presumption, however, is that this approach allows the individual patient to focus in on those concerns most relevant to him or her. It is

ABSTRACT: Most clinicians ask their patients to rate whether their health condition has improved or deteriorated over time and then use this information to guide management decisions. Many studies also use patient-rated change as an outcome measure to determine the efficacy of a particular treatment. Global rating of change (GRC) scales provide a method of obtaining this information in a manner that is quick, flexible, and efficient. As with any outcome measure, however, meaningful interpretation of results can only be undertaken with due consideration of the clinimetric properties, strengths, and weaknesses of the instrument. The purpose of this article is to summarize this information to assist appropriate interpretation of the GRC results and to provide evidence-informed advice to guide design and administration of GRC scales. These considerations are relevant and applicable to the use of GRC scales both in the clinic and in research.

KEYWORDS: Clinimetrics, Global Rating of Change, Outcome Measure

¹PhD Candidate, The George Institute for International Health, University of Sydney, Sydney, NSW, Australia

²Professor of Physiotherapy, The George Institute for International Health, University of Sydney, Sydney, NSW, Australia

³Physiotherapist, The Faculty of Health Sciences, University of Sydney, Sydney, NSW, Australia

Address all correspondence and requests for reprints to: Steven J Kamper, skamper@george.org.au

Funding: SJK and CGM are supported by the National Health and Medical Research Council of Australia.

important to note that use of the scale does not obviate the need to collect other outcome information. Rather a GRC scale may access important and relevant information additional to standardized pain and disability instruments¹³.

There is considerable variability in the title given to GRC scales cited in the literature, including Global Perceived Effect Scales¹⁹, Patient Global Impression of Change¹³, Transition Ratings¹⁷, and Global Scale¹². There is also variability in the design of the scale; for example, a recent Cochrane systematic review⁶ cited seven different designs in the eight studies that used a GRC scale. Variations may be found in the type of question asked, how many points are on the scale, and the labels assigned to the scale points. Also notable is the fact that the scales are often described in insufficient detail for readers to reproduce them. This variability may lead to confusion as to how the scale should be best used as well as difficulty in making comparisons between studies²⁰.

At present, GRC scales are widely used in research and patients are commonly asked to rate their change in condition in the clinic. Use of any method to chart outcome or clinical progress should be preceded by an understanding

of its clinimetric properties, strengths, and weaknesses. However, at this point there is no concise summary concerning GRC scales to inform clinicians and researchers as to how they can be best designed and used in a way that enhances the quality and interpretability of the information. The aims then of this paper are threefold: to provide an overview of the available clinimetric data relevant to GRC scales, to describe their strengths and weaknesses, and finally to outline recommendations for design and administration.

Clinimetric Properties

Research into the measurement properties of GRC scales is somewhat patchy in that the scale is often used as a criterion measure for the testing of other instruments, but rarely subject to rigorous investigation itself. Nevertheless, by gathering data from a cross-section of those studies that have used GRC scales, we can make some estimation of the utility of the scale. A summary of the clinimetric properties is provided in Table 1.

Face validity indicates the degree to which a measure makes sense to the reader²¹. The face validity of GRC scales is considered to be high¹⁴, and it is for this reason that these scales are fre-

quently used as a reference standard against which the validity of other outcome measures is tested²²⁻²⁴. Researchers have also measured patient ratings of the importance of a certain change concurrently with the magnitude of that change. Reported correlations for these measures are high ($r=0.72^{25}$ and $r=0.90^{26}$), and this finding also supports the face validity of a GRC indicating that gradation along the scale represents a change that is meaningful to the patient. Fischer and colleagues¹⁴ investigated the related concept of clinical relevance and reported strong correlations with patient satisfaction measures (Spearman correlation coefficients 0.56 to 0.77); these figures being significantly higher than those for serial measures.

Typically, the GRC scale has been used as an external criterion of change in studies testing domain-specific self-report measures such as pain or disability, thus providing some evidence as to the construct validity of the scale. These studies commonly present correlations between change scores on the index measure and a GRC scale. Studies have reported significant and strong correlations between GRC scores and change in measures such as the Roland Morris disability questionnaire (Pearson's $r=0.50$)²⁷, the Oswestry low back pain

TABLE 1. Summary of clinimetric properties of GRC scales.

Property	Result
Test-retest reliability	ICC 0.90 ³⁸ , 11-point
Responsiveness / Sensitivity to change	Standardized response mean 0.2–1.7 ²⁰ , 7- and 15-point Standardized response mean 0.5–2.7 ¹⁴ , 7-point
Face validity	Pearson's $r = 0.72 - 0.90$ with patient-rated importance of change ^{25,26} , 15-point
“	ICC 0.74 between clinician and patient-rated GRC ²⁶ , 15-point
“	Spearman correlation 0.87 between clinician and patient-rated change ³⁷ , 7-point
Construct validity	Significant correlation with change on Roland Morris, Oswestry, Pain rating scale, Euroqol, asthma quality of life, hop test ^{14,17,19,27,28,30} , various scales
Clinical relevance	Spearman 0.56–0.7 with patient satisfaction ¹⁴ , 7-point
MDC	0.45 points on 11-point ³⁸
MCIC	2 points on 11-point ^{9,11,38,41}
*Meaningful improvement	$\geq 5 / \leq -5$ on 15-point scale ²⁵ is meaningful improvement/deterioration ≥ 6 on 7-point scale ²⁴ is meaningful improvement

*Arbitrary designations of meaningful improvement

ICC; intraclass correlation coefficient. MDC; minimum detectable change. MCIC; minimally clinically important change.

disability questionnaire ($r=0.78$)²⁸, the pain numerical rating scale ($r=0.49$)¹⁹, the Euroqol ($r=0.42$)²⁴, and the subscales of the asthma quality of life questionnaire ($r=0.83$)¹⁷. It appears that most research using GRC for this purpose has involved the use of subjective self-report measures. However, at least one study assessed physical performance and reported a significant correlation between GRC and change in hop-test performance ($r=0.58$)²⁹ in a sample of subjects following anterior cruciate ligament repair.

Given that a patient's GRC is likely to include constructs additional to those measured by the specific measures, a perfect correlation would not be expected. Further, there is evidence to suggest that those with less severe dysfunction at baseline have smaller change scores over time³⁰⁻³²; thus, variability in baseline dysfunction levels may also reduce the strength of association between change score and GRC. What these data do indicate is the fact that patients take into account changes in pain, disability, and quality of life when assessing their global improvement or deterioration.

Of interest would be comparisons of a GRC scale with another external criterion of change. For example, some authors have tested the clinimetric properties of other instruments against criteria such as goal-attainment³⁰, return to work³³, or relevant measures of physical performance³⁴. These data would further our understanding of the construct validity of GRC and assist in interpretation of scores.

On occasion, researchers have collected clinician-rated measures of global change concurrently with patient-reported GRCs^{25,26,35-37}. Typically, an average of these scores is calculated to provide a single external change criterion. When such data are presented, however, clinician and patient-reported GRC scales appear to show acceptable agreement: 87% accuracy in classifying improved versus non-improved patients³⁶, Spearman correlation coefficient 0.87³⁷, and ICC 2,1 of 0.74²⁶.

As far as we are aware, only one study has assessed test-retest reliability. Costa et al³⁸ reported high ICC values—0.90 (95% CI 0.84 to 0.93)—indi-

cating good reproducibility in a cohort of subjects with chronic low back pain.

Fischer et al¹⁴ conducted a study aimed at directly investigating the properties of retrospective change scores on a 7-point scale and did so by comparing them to changes on repeated measures of various constructs. They measured sensitivity to change (also called *responsiveness*) by calculating the standardized response mean (change divided by standard deviation of change) and found values (0.5 to 2.7) for retrospective scores, up to three times those for serial measures. Another study²⁰ calculated the same metric using 7- and 15-point scales and reported values of 0.2 to 1.7. Sensitivity to change gives an indication of the measure's ability to capture and record a patient's change over time. Values of 0.2, 0.5, and 0.8 or greater have been designated as representing small, moderate, and large responsiveness, respectively³⁹.

The minimum detectable change of a measure gives an indication of the degree to which scores fluctuate in the absence of actual change in the patient; in effect, the measure is only capable of reliably recording changes larger than this figure. Resnik and Dobrzykowski³ outlined a method for calculating minimum detectable change (standard error of the measurement multiplied by 1.65 multiplied by $\sqrt{2}$). Using the data collected by Costa et al³⁸ from patients with chronic low back pain, this value is 0.45 points on an 11-point scale. Thus, only changes of greater than 0.45 can be considered "real" changes beyond the error inherent in the measure.

However, minimum detectable change needs to be distinguished from the minimally clinically important change, which, as the name suggests, is the change that is likely to be relevant to a patient. Norman and colleagues⁴⁰ conducted a systematic review of the size of minimally important change in quality-of-life measure and they found that such a change often corresponded to half the standard deviation for the measure. While this figure will vary according to the subject population, data from five cohorts—two acute^{9,38}, one subacute¹¹, and one chronic low back pain⁴¹, and one chronic whiplash¹⁰ using an 11-point

GRC scale—found standard deviations for the measure ranged from 1.3 to 2.7. This suggests that any change greater than 1.35 points is clinically important. Given that scores on this scale are in discrete, whole units, a change of 2 points or more on an 11-point scale may therefore be considered a clinically meaningful change.

In some studies, arbitrary cut-offs are assigned to delineate clinical importance. Van der Roer et al²⁴ collapsed *slightly better* (category 3) and *slightly worse* (category 5) subjects into the unchanged category. Stratford et al²⁵, working with a 15 point (−7 to +7) scale, defined *important improvement* as 5 or more (and deterioration as −5 or less), based on a clinical observation that patients with lesser change scores continue to seek treatment. This recommendation has been adopted subsequently by other authors^{26,42}. It should be noted that arbitrary cut-offs are generally assigned for the purposes of research and have disadvantages for determining relevant change in the clinic. Essentially, this process renders the GRC a dichotomous measure, where a particular score denotes "success" and all other scores "failure." From a clinical perspective, however, therapists may be equally as interested in progression (or deterioration) prior as they are in the achievement of a terminal end-point.

It is noted that data from a number of different scales (e.g., 7-point, 11-point, 15-point) are presented in this section. Obviously, this makes interpretation of clinimetric data problematic and urges caution in placing too much emphasis on individual indices. However, one head-to-head comparison of the measurement properties of different GRC scales has been conducted²⁰. Researchers compared a 7-point to a 15-point scale and found no significant difference in performance of the two scales in terms of their responsiveness.

Strengths of Global Rating of Change Scales

Healthcare practitioners often face stringent time constraints in the course of their practice, a fact that has implications for their choice of outcome mea-

sure. Surveys of rheumatologists in Australia⁴³ and Canada⁴⁴ found very high levels of agreement in describing the desirable qualities in outcome measures. They included simplicity, brevity, ease of scoring, reliability, validity, and sensitivity to change. Long and Dixon⁴⁵ described similar attributes in addition to that of patient relevance. Many of these qualities correspond closely to the strengths of GRC scales.

The simplicity of GRC scales makes them an attractive alternative for use in clinical practice. They are easy to administer and applicable to a wide range of patients, which is important when faced with a diverse patient population. Unlike some other generic measures designed to capture global health status, such as the SF-36 or the Sickness Impact Profile, the GRE scales are free to use, simple to score, and require no special skills or training. A GRC scale can be easily administered and interpreted which, is especially important given evidence that suggests that a very large proportion of clinicians regard time as a major barrier to the use of formal outcome measures⁴⁶.

The concept of clinical relevance has been identified as critical in choice of outcome measure^{47,48}. The nature of the question asked can ensure that a GRC scale satisfies this criterion. An open question leaves the patient to decide what construct(s) he or she considers important in determining health status; hence, clinical relevance to the individual is assured. The validity of such subjective self-ratings of health has also been demonstrated⁴⁹ with correlation between these measures and various biological markers of health and mortality in older people. Correlations with self-rated importance of change^{25,26} and patient satisfaction measures¹⁴ described in the previous section also attest to the clinical relevance of GRC scales.

GRC scales must be adapted to suit the needs of the clinician. By tailoring the question, the scale is made relevant to whatever health condition a patient presents with. This offers the advantages of a condition-specific measure for any

condition. The scale must also be adapted to measure the time period of interest; for example, the clinician or researcher might be interested in change from the onset of a condition or, conversely, change since the last contact.

Weaknesses of Global Rating of Change Scales

A notable criticism of GRC scales is founded in Ross's theory of implicit change¹⁶ and relates to the way in which people construct their memories. In brief, this theory states that people are unlikely to accurately recall a previous state or attitude; rather they create an impression of how much they have changed by considering their present condition and then retrospectively applying some idea of their change over time. This is wholly different from the simple subtraction of one measurement from another, which is the implicit function of GRC scales. Ross contended that people formulate a view regarding the presence or absence of change in a certain construct and base their change score on this view. This may lead to either understatement or exaggeration in the score as compared to the actual change as calculated by the difference between serial measurements.

Other researchers have also questioned the reliability of patients' estimates of previous health status. Hermann¹⁵ described the problem of "recall bias" where events intervening between the anchor points influence the recall of the original status, and Schwartz and Sprangers⁵⁰ described "response-shift" where a patient's response is influenced by a changing perception of their context. The questionable ability of patients to accurately recall and score a previous health state underpins the principal criticism of GRC scales, which is that scores are unduly influenced by a patient's current status rather than measuring transition as intended. Indeed, empirical evidence of this criticism has been reported in studies that show that the correlation between a GRC score and the current measure on another instrument is often stronger than the cor-

relation between GRC and the change between serial measures on that instrument^{17,18}.

There is some uncertainty regarding the reproducibility of GRC scales, i.e., the likelihood that scores will remain stable in the event that no change occurs in the patient. This uncertainty is founded in the contention that single-item measures are less reliable than multi-item¹². Although this is not always supported by empirical evidence⁵¹, the issue is well summarized by Sloan and colleagues⁵². There are some data, however, as presented in the previous section, that suggest that reproducibility of GRC scales may be adequate³⁸.

The fact that GRC scales measure whatever construct the patient chooses as most appropriate is at once a great strength and also a weakness of the measure. The concept of recovery is complex and most likely multidimensional⁵³, and the assessor does not know what it is the patient takes into account when making the rating, for example, pain, functional limitations, quality of life, side effects⁵⁴, etc. Further, it is likely that successive patients will consider different aspects of their health to be important and hence respond based on a different set of parameters. This makes assessing the construct validity of GRC scales problematic and also demands that any such "global" assessment be considered in context with other measures.

Design and Administration of the Global Rating of Change Scale

The function of GRC scales is dependent on the two constituent parts⁵⁵: first, the question that is asked of the patient and second, the scale upon which the response is scored. The following section contains discussion relevant to the design of both of these. Where possible, design recommendations are based on empirical evidence and references provided; however, where such data was not available, the authors provide opinion based on their clinical and research experience. The section describing formulation of the question consists largely of opinion-based recommendations; this

is necessitated by the lack of research that directly investigates the differences between scales.

Formulating the Question

In order to avoid ambiguity and ensure that the patient provides information relevant to the condition in which the assessor is interested, it is important that the health condition is mentioned explicitly in the question. This is particularly important if the patient has comorbidities that are not the target of the intervention. For example, a patient being treated for a knee injury may also have a chronic respiratory condition. If the clinician is interested in the patient's recovery from the knee injury, then reference to it should be included in the question, e.g., "With respect to your knee injury . . ."

The wording of the question will direct the patient towards the construct that the scale will measure. This means that the clinician must decide what it is that he or she is looking to measure. For example, the assessor may be interested in patient satisfaction with outcome, self-assessed health status, or quality of life. In each case, the question will be

worded differently. Whatever construct the assessor is interested in, the question is left open so as to allow the patient to decide what he or she will take into account in determining the response. This ensures clinical relevance, for example, "With respect to your low back pain, how satisfied are you with your condition . . ." or "With respect to your ankle sprain, how would you describe yourself . . ."

The final step in creating a question is to provide an anchor for the scale. This is the previous time point to which the patient's current status will be compared. It is likely that choosing an anchor that corresponds to a significant event will improve the ability of a patient to recall health status at that time¹⁷ and therefore optimize the reliability of the score. This event might be an accident, surgery, acute onset of a condition, or the commencement of treatment. For example, "With respect to your wrist injury, how would you describe yourself now compared to when your cast was removed?"

Addressing each of the criteria outlined above minimizes ambiguity so that patient response is not biased by error associated with misunderstanding the question. An example of an appropriate

question to measure change in global health status is presented in Figure 1.

Designing the Scale Itself

The patient responds to the question by marking a scale that measures the degree of improvement or deterioration. When designing this scale, the assessor must first consider how many points or response options there should be. This number may range from the simplest 3-point scale⁵⁶ to a 101-point scale⁵.

Having too few points risks losing information. For example, in the case of a 3-point scale (better, the same, worse), there would be no difference in the score if a patient improved just marginally as opposed to completely recovered. Assuming that a patient can discriminate between varying degrees of improvement or deterioration, the scale will not be sensitive enough to delineate. There is also empirical evidence that scales with such low numbers of response items are less reliable and valid⁵⁷. Conversely, there may also be problems associated with having too many response categories, although the effect is less clear. While it may be that having a large number, for example, 101, does not ad-

"With respect to your whiplash injury, how would you describe yourself now compared to immediately after your accident?"

FIGURE 1. Recommended GRC scale.

versely affect measurement properties⁵⁷, there is the concern that patients will have difficulty in attaching meaning to individual points, which leads to reduced consistency in responses⁵⁸.

Preston and Colman⁵⁷ measured what patients thought of scales with different numbers of response options and found that they preferred those with 7, 9, or 10 categories. This finding is consistent with Miller's work on human information-processing capacity⁵⁹. Overall, scales with 7 to 11 points appear to offer the best compromise between patient preference, adequate discriminative ability, and test-retest reliability⁵⁷. This range also corresponds to the number favored most commonly in the research literature^{19,60,61}. A characteristic of rating scales is that patients tend not to respond at the extreme ends of a scale, in so-called end-aversion bias⁶². While it is not possible to avoid this problem altogether, it can be taken into account when choosing the number of scale points. With this in mind, we suggest that it may be prudent to choose a number of scale points toward the higher end of the optimal range, on the understanding that some of the end points will rarely be used.

GRC scales used in published research vary as to whether the scale points have numerical and/or written descriptors attached to them. In some cases, numerical labels include negative numbers¹⁹ and others do not⁶¹. Written descriptors may be placed only on end points⁷, on all points⁸, or on not included at all⁶⁰.

The ideal scale should be balanced, with equal numbers of points on either side of a midpoint labelled *unchanged*. The midpoint is important so that the patient is not forced to rate himself or herself as better or worse in the event that the condition has remained the same⁵⁸. Endpoints should also be labelled with written descriptors in order to assign meaning to the scale⁵⁸. These descriptors also need to be congruent with the question. Anchoring the positive end of the scale with *completely recovered* or some variant, supplies the scale with a terminal end point.

Placing numeric descriptors on the other points is preferable to written ones

as this makes the scale less cluttered and avoids possible ambiguity in the language used⁶³. Since GRC scales are designed to measure a bipolar construct, i.e., improvement versus deterioration, and in order then to reinforce the intended meaning of the scale points, it would seem appropriate to have a scale balanced around zero (0) (unchanged) with negative (deterioration) and positive (improvement) numbers⁵⁸.

Administration of the Measure

When using a GRC scale to measure outcome, the mode of administration should be standardized. Having a patient take the measure away and complete it in private as opposed to the assessor asking the question and marking the scale may alter the way the patient scores the scale⁶⁴. In clinical practice, where the instrument is used to measure within-patient change, administration should be standardized for each assessment occasion.

How the passage of time affects the measurement properties of the measure is largely unknown. Intuitively, one might expect that a longer period of transition would adversely impact on a patient's ability to recall a prior health state and hence lead to less reliable information. Although some studies report on follow-up periods of 4–6 weeks^{14,17}, there is to date no empirical evidence to support a recommendation on the maximum length of time over which a change measure should be scored. For the purposes of clinical practice where patients are likely to be seen on regular occasions, at least fortnightly, this is unlikely to represent a significant issue.

Summary

GRC scales offer a flexible, quick, and simple method of charting self-assessed clinical progress in research and clinical settings. The instrument has the advantages of clinical relevance, adequate reproducibility, and sensitivity to change and is intuitively easy to understand by the patient and the person administering. While scores correlate with pain, disability, and quality-of-life measures,

the open nature of the question allows the patient to take into account other factors that he or she may consider important in his or her clinical situation. It is likely, however, that a patient's condition at the time of asking exerts a significant influence over the GRC score and the clinician/researcher should consider this when interpreting results. Practically speaking, this may mean that patients with lower symptom severity score a greater positive change on the GRC and vice-versa. Additionally, it should be recognized that increasing length of recall time may adversely affect the validity of GRC scores. It may be that evaluation of serial measures is preferable to self-reported change when transition time stretches into several months.

Measurement of clinical outcome generally, and of complex subjective constructs such as recovery in particular, is an inherently difficult process. The choice of instrument necessarily requires balancing a variety of practical and clinimetric concerns. In recognizing that no single instrument provides the answer as far as outcome goes, it is important then to consider data from any outcome measure not in isolation but within the wider clinical context.

While there is considerable variability in the way in which GRC scales are designed and administered, interpretability and reliability can be optimized by paying attention to a number of design features. The authors recommend explicit mention of the specific condition, construct, and time anchor point in the question. Evidence supports the use of a balanced 7–11 point numerical scale with written descriptors on the ends and at the midpoint. Administration mode should be standardized.

REFERENCES

1. Beaton DE. Understanding the relevance of measured change through studies of responsiveness. *Spine* 2000;25:3192–3199.
2. American Physical Therapy Association. Guide to Physical Therapy Practice. 2nd ed. *Phys Ther* 2001;81:9–746.
3. Resnik L, Dobrzykowski E. Guide to outcome measurement for patients with low back pain syndromes. *J Orthop Sports Phys Ther* 2003;33:307–318.

4. Gridley L, van den Dolder PA. The percentage improvement in pain scale as a measure of physiotherapy treatment effects. *Aust J Physiother* 2001;47:133-136.
5. Jull G, Trott P, Potter H, et al. A randomized controlled trial of exercise and manipulative therapy for cervicogenic headache. *Spine* 2002;27:1835-1843.
6. Kay TM, Gross A, Goldsmith C, Santaguida PL, Hoving J, Bronfort G. Exercises for mechanical neck disorders. *Cochrane Database Syst Rev* 2005;CD004250.
7. Koes BWM, Bouter LMP, van Mameren HP, et al. The effectiveness of manual therapy, physiotherapy, and treatment by the general practitioner for nonspecific back and neck complaints: A randomized clinical trial. *Spine* 1992;17:28-35.
8. Pennie BH, Agambar LJ. Whiplash injuries: A trial of early management. *J Bone Joint Surg Br* 1990;72-B:277-279.
9. Hancock MJ, Maher CG, Latimer J, et al. Assessment of diclofenac or spinal manipulative therapy, or both, in addition to recommended first-line treatment for acute low back pain: A randomised controlled trial. *Lancet* 2007;370:1638-1643.
10. Stewart MJ, Maher CG, Refshauge KM, Herbert RD, Bogduk N, Nicholas M. Randomized controlled trial of exercise for chronic whiplash-associated disorders. *Pain* 2007;128:59-68.
11. Pengel LHM, Refshauge KM, Maher CG, et al. Nicholas M, Herbert RD, McNair P. Physiotherapist-directed exercise, advice, or both for subacute low back pain: A randomized trial. *Ann Intern Med* 2007;146:787-796.
12. 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.
13. Dworkin RH, Turk DC, Farrar JT, et al. Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. *Pain* 2005;113:9-19.
14. Fischer D, Stewart AL, Bloch DA, Lorig K, Laurent D, Holman H. Capturing the patient's view of change as a clinical outcome measure. *JAMA* 1999;282:1157-1162.
15. Herrmann D. Reporting current, past and changed health status: What we know about distortion. *Med Care* 1995;33:AS89-AS94.
16. Ross M. Relation of implicit theories to the construction of personal histories. *Psychol Rev* 1989;96:341-357.
17. Guyatt GH, Norman GR, Juniper EF, Griffith LE. A critical look at transition ratings. *J Clin Epidemiol* 2002;55:900-908.
18. Schmitt J, Di Fabio RP. The validity of prospective and retrospective global change criterion measures. *Arch Phys Med Rehabil* 2005;86:2270-2276.
19. Stewart M, Maher CG, Refshauge KM, Bogduk N, Nicholas M. Responsiveness of pain and disability measures for chronic whiplash. *Spine* 2007;32:580-585.
20. Lauridsen HH, Hartvigsen J, Korsholm L, Grunnet-Nilsson N, Manniche C. Choice of external criteria in back pain research: Does it matter? Recommendations based on analysis of responsiveness. *Pain* 2007;131:112-120.
21. Dijkers M. Measuring quality of life: Methodological issues. *Am J Phys Med Rehabil* 1999;78:286-300.
22. Rebbeck TJ, Refshauge KM, Maher CG, Stewart M. Evaluation of the core outcome measure in whiplash. *Spine* 2007;32:696-702.
23. Scrimshaw SV, Maher CG. Responsiveness of visual analogue and McGill pain scale measures. *J Manipulative Physiol Ther* 2001;24:501-504.
24. van den Roer N, Ostelo RWJG, Bekkering GE, van Tulder MW, De Vet HCW. Minimal clinically important change for pain intensity, functional status, general health status in patients with nonspecific low back pain. *Spine* 2006;31:578-582.
25. Stratford PW, Binkley JM, Solomon P, Gill C, Finch E. Assessing change over time in patients with low back pain. *Phys Ther* 1994;74:528-533.
26. Watson CJ, Propps M, Ratner J, Zeigler DL, Horton P, Smith SS. Reliability and responsiveness of the lower extremity functional scale and the anterior knee pain scale in patients with anterior knee pain. *J Orthop Sports Phys Ther* 2005;35:136-146.
27. Pengel LHM, Refshauge KM, Maher CG. Responsiveness of pain, disability and physical impairment outcomes in patients with low back pain. *Spine* 2004;29:879-885.
28. Fritz JM, Irrgang JJ. A comparison of a modified Oswestry low back pain disability questionnaire and the Quebec back disability scale. *Phys Ther* 2001;81:776-788.
29. Reid A, Birmingham TB, Stratford PW, Alcock GK, Giffin JR. Hop testing provides a reliable and valid outcome measure during rehabilitation after anterior cruciate ligament reconstruction. *Phys Ther* 2007;87:337-349.
30. Riddle DL, Stratford PW, Binkley JM. Sensitivity to change of the Roland-Morris Back Pain Questionnaire: Part 2. *Phys Ther* 1998;78:1197-1207.
31. Stratford PW, Binkley JM, Riddle DL, Guyatt GH. Sensitivity to change of the Roland-Morris Back Pain Questionnaire: Part 1. *Phys Ther* 1998;78:1186-1196.
32. Stucki G, Daltroy L, Katz JN, Johannesson M, Liang MH. Interpretation of change scores in ordinal clinical scales and health status measures: The whole may not equal the sum of the parts. *J Clin Epidemiol* 1996;49:711-717.
33. Kvale A, Skoun JS, Ljunggren AE. Sensitivity to change and responsiveness of the Global Physiotherapy Examination (GPE-52) in patients with long-lasting musculoskeletal pain. *Phys Ther* 2005;85:712-726.
34. Gronblad M, Jarvinen E, Hurri H, Hupli M, Karaharju E. Relationship of the Pain Disability Index (PDI) and the Oswestry disability questionnaire (ODQ) with three dynamic physical tests in a group of patients with chronic low-back and leg pain. *Clin J Pain* 1994;10:197-203.
35. Deyo RA, Inui TS. Toward clinical applications of health status measures: Sensitivity of scale to clinically important changes. *Health Serv Res* 1984;19:275-289.
36. Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: An analogy to diagnostic test performance. *J Chronic Diseases* 1986;39:897-906.
37. Farrar JT, Young JP, LaMoreaux L, Werth JL, Poole RM. Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. *Pain* 2001;94:149-158.
38. Costa LOP, Maher CG, Latimer J, et al. Clinimetric testing of three self-report outcome measures for low back pain patients in Brazil: Which one is the best? *Spine* 2008;33:2459-2463.
39. Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing responsiveness: A critical review and recommendations. *J Clin Epidemiol* 2000;53:459-468.
40. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: The remarkable universality of half a standard deviation. *Med Care* 2003;41:582-592.
41. Ferreira ML, Ferreira PH, Latimer J, et al.

- Comparison of general exercise, motor control exercise and spinal manipulative therapy for chronic low back pain: A randomized trial. *Pain* 2007;131:31–37.
42. Perillo M, Bulbulian R. Responsiveness of the Bounremouth and Oswestry questionnaires: A prospective pilot study. *J Manipulative Physiol Ther* 2003;26:77–86.
 43. Bellamy N, Muirden KD, Brooks PM, Barracough D, Tellus MM, Campbell J. A survey of outcome measurement procedures in routine rheumatology outpatient practice in Australia. *J Rheumatol* 1999;26:1593–1599.
 44. Bellamy N, Kaloni S, Pope J, Coulter K, Campbell J. Quantitative rheumatology: A survey of outcome measurement procedures in routine rheumatology outpatient practice in Canada. *J Rheumatol* 1998;25:852–859.
 45. Long AF, Dixon P. Monitoring outcomes in routine practice: Defining appropriate measurement criteria. *J Eval Clin Pract* 1996;2:71–78.
 46. Abrams D, Davidson M, Harrick J, Harcourt P, Zylinski M, Clancy J. Monitoring the change: Current trends in outcome measurement usage in physiotherapy. *Man Ther* 2006;11:46–53.
 47. Bombadier C. Outcomes assessment in the evaluation of treatment of spinal disorders. *Spine* 2000;25:3100–3103.
 48. Middel B, Stewart R, Bouma J, van Sonderen E, van den Heuvel WJA. How to validate clinically important changes in health-related functional status: Is the magnitude of effect size consistently related to magnitude of change as indicated by a global question rating? *J Eval Clin Pract* 2001;7:399–410.
 49. Jylha M, Volpato S, Guralnik JM. Self-rated health showed a graded association with frequently used biomarkers in a large population sample. *J Clin Epidemiol* 2006;59:465–471.
 50. Schwartz CE, Sprangers MAG. Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. *Soc Sci Med* 1999;48:1531–1548.
 51. Gardner DG, Cummings LL, Dunham RB, Pierce JL. Single-item versus multiple-item measurement scales: An empirical comparison. *Ed & Psych Measurement* 1998;58:898–915.
 52. Sloan JA, Aaronson N, Cappelleri JC, Fairclough DL, Varricchio C. Assessing the clinical significance of single items relative to summated scores. *Mayo Clinic Proceed* 2002;77:479–487.
 53. Beaton DE, Tarasuk V, Katz JN, Wright JG, Bombadier C. Are you better? A qualitative study of the meaning of recovery. *Arthritis Care & Res* 2001;45:270–279.
 54. Dworkin RH, Turk DC, Wyrwich KW, et al. Interpreting the clinical importance of treatment outcomes in chronic pain trials: IMMPACT recommendations. *J Pain* 2008;9:105–121.
 55. Hudak PL, Wright JG. The characteristics of patient satisfaction measures. *Spine* 2000;25:3167–3177.
 56. Jaeschke R, Singer J, Guyatt GH. Measurement of health status: Ascertaining the minimal clinically important difference. *Controlled Clinical Trials* 1989;10:407–415.
 57. Preston CC, Colman AM. Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psych* 2000;104:1–15.
 58. Krosnick JA, Fabrigar LR. Designing rating scales for effective measurement in surveys. In Lyberg, Biemer, Collins, et al, eds. *Survey Measurement and Process Quality*. Hoboken, NJ: John Wiley & Sons, 1997.
 59. Miller GA. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psych Rev* 1956;63:81–97.
 60. Haspeslagh S, van Suijlekom H, Lame I, Kessels A, van Kleef M, Weber W. Randomised controlled trial of cervical radio-frequency lesions as a treatment for cervicogenic headache. *BMC Anesthes* 2006;6:1.
 61. Ostelo RWJG, de Vet HCW. Clinically important outcomes in low back pain. *Best Pract Res Clin Rheum* 2005;19:593–607.
 62. Streiner DL, Norman GR. *Health Measurement Scales: A Practical Guide to Their Development and Use*. 3rd ed. New York: Oxford University Press, 2003.
 63. American Statistical Association. The influence of administration mode on responses to numeric rating scales. Annual Meeting of the American Statistical Association. 1994; Alexandria, VA.
 64. Tourangeau R, Smith TW. Asking sensitive questions: The impact of data collection mode, question format and question context. *Public Opinion Quarterly* 1996;60:275–304.