Research Report

Defining the Minimum Level of Detectable Change for the Roland-Morris Questionnaire

Background and Purpose. The Roland-Morris Questionnaire (RMQ) is a self-administered disability measure in which greater levels of disability are reflected by higher numbers on a 24-point scale. The RMQ has been shown to yield reliable measurements, which are valid for inferring the level of disability, and to be sensitive to change over time for groups of patients with low back pain. Little is known about the usefulness of this instrument in aiding decision making regarding individual patients. The purpose of this study was to determine the minimum level of detectable change when the RMQ is applied to individual patients. Subjects. The study sample consisted of 60 outpatients with low back pain. Methods. The RMQ was administered at the subjects' initial visit and again 4 to 6 weeks later. Conditional standard errors of measurement (CSEMs) were computed for initial and follow-up RMQ scores, and these values were used to estimate the minimum level of detectable change. Results. Minimum levels of detectable change at the 90% confidence level varied from 4 to 5 RMQ points. Conclusion and Discussion. The magnitude of CSEMs is sufficiently small to detect change in patients with initial scores in the central portion of the scale (4–20 RMQ points); however, the magnitude is too large to detect improvement in patients with scores of less than 4 and deterioration in patients who have scores greater than 20. [Stratford PW, Binkley J, Solomon P, et al. Defining the minimum level of detectable change for the Roland-Morris Questionnaire. *Phys Ther.* 1996;76:359–365.]

Key Words: *Backache, Disability, Evaluation, Outcome.*

Paul W Stratford

Jill Binkley

Patricia Solomon

Elspeth Finch

Caroline Gill

Julie Moreland

Physical therapists regularly use measurements (eg, of range of motion, pain, or disability) to determine whether a patient's status has changed over time. Often, when the results differ from one assessment to the next, therapists assume patients have undergone true change. Unfortunately, some or all of the difference between assessments can occur as a result of measurement error, including random fluctuation in patients who may or may not have truly changed. A patient who at the initial assessment scores 14 out of a possible 24 points on a particular disability questionnaire and 4 weeks later scores 10 points may appear to have undergone change. The 4-point difference may represent true change, or it could fall within the limits of measurement error and inherent variability in a truly unchanged patient. The importance of this issue is underscored when the management options available to therapists are considered. For example, therapists who view the difference in scores as representing true change may elect to continue with an intervention. Therapists who consider the 4-point change to be within the limits of measurement error, however, may choose to alter the intervention in hopes of selecting a treatment that is more effective. The goal of this report is to provide clinicians with guidelines for assessing change over time when they use the Roland-Morris Questionnaire (RMQ)[1,2] to assess disability in patients with low back pain (LBP).

The RMQ was selected because its measurement properties have been shown to be equal to or better than those of similar measures used to assess change in disability in patients with LBP.[1–15] In Table 1, we provide a brief summary of the more frequently used and researched measures. The RMQ is a self-administered questionnaire consisting of 24 items chosen from the Sickness Impact Profile (SIP).[16] Items were chosen to reflect a variety of activities of daily living. To improve the specificity of the response, Roland and Morris[1] added the phrase "because of my back" to each item. An item receives a score of 1 if it is checked as applicable by the respondent and a score of 0 if it is not marked. Accordingly, total scores can vary from 0 (no disability) to 24 (severe disability). Research of the RMQ's measurement properties has provided consistent estimates of internal consistency, test-retest reliability (accounting for the interval between assessments), construct validity, and sensitivity-to-change coefficients. The term "sensitivity to change" describes a measure's ability to assess change over time.

One strategy for assessing and reporting change over time, reported in Table 1, is the receiver operating characteristic (ROC) curve.[15] With this technique, sensitivity (y-axis) is plotted against 1-specificity (x-axis). When assessing change over time, *sensitivity* is defined as the number of patients correctly identified (by a given questionnaire) as having undergone a clinically important change divided by all patients who truly underwent a clinically important change. *Specificity* refers to the number of patients who were correctly identified (by a given questionnaire) as not undergoing a clinically important change divided by all patients who truly did not undergo a clinically important change. The greater the area under the curve, the greater a questionnaire's ability to distinguish patients who did and did not undergo a clinically important change. The area under the curve can be interpreted as the probability of correctly identifying a patient who has undergone a clinically important change from randomly selected pairs of patients who have and have not undergone an important change.

PW Stratford, MSc, PT, is Assistant Professor, Faculty of Health Sciences, School of Rehabilitation Science, McMaster University, Bldg T16, 1280 Main St W, Hamilton, Ontario, Canada L8S 4K1 (stratfor@mcmaster.ca). Address all correspondence to Mr Stratford.

J Binkley, MClSc, PT, COMP, is Director of Research, Rehab Management Systems, Dahlonega, GA 30597. She was an orthopedic clinical specialist and Assistant Professor, Department of Physical Therapy, North Georgia College, Dahlonega, GA 30597, at the time of this study.

P Solomon, PhD, PT, is Assistant Professor, Faculty of Health Sciences, School of Rehabilitation Science, McMaster University.

E Finch, MHSc, PT, is Assistant Professor, Faculty of Health Sciences, School of Rehabilitation Science, McMaster University.

C Gill, PT, is Senior Physiotherapist–Orthopaedics, St Joseph s Hospital, Hamilton, Ontario, Canada.

J Moreland, MSc, PT, is Research Therapist, St Joseph's Hospital and St Peter's Hospital, and Assistant Clinical Professor, Faculty of Health Sciences, School of Rehabilitation Science, McMaster University.

This study was approved by the Ethics Committee of St Joseph's Hospital.

This article was submitted February 9, 1995, and was accepted November 28, 1995.

Table 1.
Summary of Measurement Properties of Several Measures Used to Assess Patients With Low Back Pain[a]

	Roland-Morris[1]	Oswestry[3]	Waddell[4]	SF-36[5]
Reliability				
Internal Consistency	α=.90[6] α=.84[7] α=.89, .92[8]	α=.87[6] α=.77, .93[8]	α=.76[4]	α=.89[6]
Test-retest	ICC=.91 (<2 wk)[6] ICC=.86 (3–6 wk)[9] r=.83 (3 wk)[10]	ICC=.91 (<2 wk)[6] ICC=.83 (1 wk)[11]		ICC=.65 (<2 wk)[6]
Validity	Quebec, r=.77[6] Pain, r=.38[11] Pain, r=.41[10] SIP, r=.85[10]	Quebec, r=.80[6] Pain, r=.47[11]	OSW, r=.70[4]	Quebec, r=.72[6] Back pain scale, r=.36–.69[12] Patient generated index, r=.18–.47[13]
Sensitivity to change over time				
Change only	SRM=.50[6]	SRM=.36[6]		SRM=.15[6] SRM=.09–.50[12]
Validity of change	Global rating, r=.47[6] Global rating, r=.60[14] Oswestry, r=.79[14] ROC area=.72[15] ROC area=.79[14]	Global rating, r=.35[6] Global rating, r=.57[14] RMQ, r=.79[14] ROC area=.78[14]	Able to discriminate success and failure in patients with acute attack of LBP[4]	Global rating, r=.31[6]

[a] ICC=intraclass correlation coefficient, OSW=Oswestry Low Back Pain Disability Questionnaire, SIP=Sickness Impact Profile, SRM=standardized response mean, RMQ=Roland-Morris Questionnaire, ROC=receiver operating characteristic, LBP=low back pain.

Numerous studies[4,6,9,10,12,14,15] have assessed measures of sensitivity to change in patients with LBP. In only one of the many investigations reported, however, was information presented in a format that is suitable for making decisions on individual patients.[9] Using a test-retest reliability design (the interval between assessments was 3–6 weeks), Stratford et al[9] calculated the standard error of measurement (SEM) for RMQ scores in 36 stable patients with LBP to be 1.79 RMQ points. The SEM expresses measurement error in the same units as those of the original measurement, in this case RMQ points. The SEM is a measure of within-patient variability and is calculated by taking the square root of the mean square error term from the usual reliability study analysis-of-variance table. In addition to reporting the SEM, these authors calculated the minimal level of detectable change at the 95% confidence level to be 5 RMQ points.[9] This value defines the smallest difference that can be detected between two measurements. It is also referred to as the "reliability change index."[17] The interpretation of the minimal level of detectable change is that an observed change in a patient that is less than this value is deemed to be indistinguishable from measurement error. Accordingly, a patient who demonstrates a change score that is less than this value is viewed as not having undergone change. The principal limitation of early work reporting the SEM[9] is that this statistic assumes measurement error is constant across the range of possible scores. In this report, the conditional standard error of measurement (CSEM) will be used. This measure is defined in the "Method" section, and an illustration is provided in the Appendix.

In this study, we attempted to improve on previous research by providing clinicians with estimates of minimal detectable change using a process that takes into account that the level will change for different combinations of initial and follow-up RMQ score comparisons. These estimates can be used to determine whether the disability of an individual patient is likely to have actually changed. The research question was: What are the minimum levels of detectable change for all possible score comparisons on the RMQ when it is applied to patients with LBP?

Method

Subjects

The sample consisted of 60 patients with LBP (37 male, 23 female) who were referred by their physicians to the outpatient physical therapy departments of two hospitals. Patients were eligible for this study if they (1) were diagnosed as having LBP of apparent musculoskeletal origin, (2) could read English, and (3) provided written consent on a form approved by the centers' research review boards. The patients were aged 18 to 72 years ($\bar{X}$=41, SD=12). Forty-eight patients were employed at the time of onset of the episode of back pain associated

with these referrals, and the referrals of 35 of these patients involved insurance claims. Thirty-eight patients experienced sudden onset of discomfort, 20 patients experienced a gradual onset of discomfort, and 2 patients were uncertain as to the nature of the onset of discomfort. Nineteen patients had a limited straight leg raise (estimated at less than 60°), and 12 patients had episodes of discomfort distal to the knee at the time of initial assessment. This episode of LBP was less than 6 weeks for all patients. The sample size of 60 patients was based on an expected internal consistency coefficient of .90[6–8] and a lower 95% confidence interval width of .05.[23]

Design

A before-after study design was used to obtain two RMQ scores for each patient. Patients completed the RMQ prior to beginning physical therapy and following 4 to 6 weeks of treatment. This duration was chosen for two reasons: (1) The natural history of acute LBP is such that over 60% of patients show significant improvement within this interval,[2] and (2) intervention studies on patients with acute LBP often report outcomes between 4 and 6 weeks.[18,19] It is important to note that the physical therapy interventions applied to patients were neither of interest nor under investigation. These interventions, like the interval between assessments, served as a construct for achieving a change. Patients were asked to complete the RMQ in accordance with the instructions provided by Roland and Morris.[1] This process allowed estimates of measurement error to be assessed for both points in time.

Data Analysis

Conditional standard errors of measurement were used to estimate the minimum levels of detectable change.[20] The method is based on the binomial theory of measurement error[21] and the correction approach described by Keats.[22*] All possible initial and follow-up score combinations were compared using the Z statistic. A 90% confidence level was chosen, and the corresponding Z value is 1.65. An illustration of the analysis is provided in the Appendix. Actual patient data were required only to estimate the reliability coefficients used for this correction factor.

*In brief, the binomial theory of measurement error dictates that when item scoring is dichotomous, as it is on the RMQ, the error variance (σ_E^2) for any given score is equal to

$$\left[\frac{(n - X_p)(X_p)}{n - 1}\right]$$

where n equals the number of items on the test and X_p is the patient's RMQ score. The Keats correction factor takes into account that the variance formula tends to overestimate σ_E^2 when the forms are similar.[22] To apply the Keats correction factor, σ_E^2 is multiplied by

$$\left[\frac{1 - KR_{20}}{1 - KR_{21}}\right]$$

where KR_{20} and KR_{21} are the Kuder-Richerson reliability coefficients.[23] The CSEMs were determined for all possible RMQ scores.

Results

The means and 90% confidence intervals for the initial and follow-up RMQ scores were 11.5 (9.9–13.1) and 6.6 (5.1– 8.1) RMQ points, respectively. The KR_{20} reliability coefficient was .92 for both the initial and follow-up visits, whereas the KR_{21} coefficients for the initial and follow-up visits were .89 and .90, respectively. Table 2 provides a summary of the initial and follow-up conditional error variances and CSEMs for all possible RMQ scores. For example, the CSEM for initial and follow-up RMQ scores of 14 are 2.13 and 2.24, respectively. The small difference in CSEM scores between initial and follow-up visits of the same score is due to the slight difference in the magnitude of the KR_{21} coefficient for these two points in time. The Figure provides a summary of minimum detectable change values for improvement and deterioration. The data points for this figure were calculated in accordance with the procedure outlined in step 6 of the Appendix. In order to be confident at the 90% level that a change has occurred, the intersection of the initial and follow-up scores must be outside the shaded area. For example, a patient who had an initial score of 14 must achieve a score of 9 or lower for the clinician to be confident that improvement has occurred, or a score of 18 or higher to be convinced that deterioration has taken place. Finally, the Figure shows that improvement cannot be detected for patients who have initial RMQ scores lower than 4 and that deterioration cannot be ascertained for patients who have initial RMQ scores greater than or equal to 20.

Discussion

Researchers using the SEM have estimated the minimum level of detectable change to be approximately 5 RMQ points at the 95% confidence level.[9] A limitation of using the SEM is that it assumes that the magnitude of measurement error is uniform across the entire scale (ie, equal for all scores). Moreover, a shortcoming of our previous study[9] was that most of the patients' initial RMQ scores were in the central portion of the scale. Accordingly, a clinician cannot be confident that a change of 5 RMQ points accurately reflects the minimum level of detectable change for values more distant than those located near the central portion of the scale. One strategy for estimating the score-specific level of minimum detectable change would be to conduct a number of reliability studies in which patients are stratified on the basis of their initial scores. To obtain a reasonable confidence interval on the reliability coefficient, approximately 30 stable patients per group would be required. Given that patients can have 25 possible initial RMQ scores (ie, 0–24), 25 studies would be required. The

Table 2.
Error Variances and Conditional Standard Errors of Measurement (CSEMs) for Various Roland-Morris Questionnaire Scores

Roland-Morris Questionnaire Score	Initial Visit Error Variance	Initial Visit CSEM	Follow-up Error Variance	Follow-up CSEM
0	0.38	0.61	0.42	0.64
1	0.75	0.86	0.83	0.91
2	1.43	1.19	1.58	1.26
3	2.04	1.43	2.26	1.50
4	2.59	1.61	2.87	1.69
5	3.08	1.75	3.41	1.85
6	3.50	1.87	3.88	1.97
7	3.86	1.96	4.27	2.07
8	4.15	2.04	4.59	2.14
9	4.38	2.09	4.84	2.20
10	4.54	2.13	5.02	2.24
11	4.64	2.15	5.13	2.27
12	4.67	2.16	5.17	2.27
13	4.64	2.15	5.13	2.27
14	4.54	2.13	5.02	2.24
15	4.38	2.06	4.84	2.20
16	4.15	2.04	4.59	2.14
17	3.86	1.96	4.27	2.07
18	3.50	1.87	3.88	1.97
19	3.08	1.75	3.41	1.85
20	2.59	1.61	2.87	1.69
21	2.04	1.43	2.26	1.50
22	1.43	1.19	1.58	1.26
23	0.75	0.86	0.83	0.91
24	0.38	0.61	0.42	0.64

feasibility of such a venture, however, due to costs and the availability of patients, is unlikely.

In our study, we attempted to address the deficiency of previous work by calculating the CSEM and the minimum level of detectable change for various initial and follow-up score combinations. Rather than performing a stratified study, we estimated the CSEM using the binomial theory of measurement error. Using this approach, the magnitude of measurement error is dependent on the actual scores being compared. This approach is appropriate when item scoring is dichotomous, as in the RMQ. Our results are consistent with those of previous work to the extent that initial scores located near the central portion of the scale require a change of 5 RMQ points for a clinician to be confident at the 95% level that a change has really occurred. The results, however, add to existing knowledge by suggesting that a change of only 4 RMQ points is required to detect improvement in patients with initial scores of 4 to 11 RMQ points and in patients with scores greater than 16 RMQ points. Similarly, a change of only 4 RMQ points is needed to detect deterioration in patients with initial scores lower than 7 RMQ points and in patients with scores of 13 to 20 RMQ points. Improvement in patients with initial RMQ scores lower than 4 RMQ points and deterioration in patients with initial scores greater than 20 RMQ points cannot be detected at the 90% level, and decisions for such patients must be made at a lower confidence level.

We believe our findings can be generalized to various clinical settings. The ages, gender distribution, and initial and follow-up RMQ scores of our sample are consistent with those of other researchers reporting on patients with acute or subacute LBP.[9,14,15,18,19,24,25] Furthermore, the magnitudes of the KR_{20} and KR_{21} coefficients display a remarkable similarity to internal consistency coefficients for the RMQ reported by other authors.[6–8] For these reasons, we believe that the reported levels of minimum detectable change are generalizable to patients with acute and subacute LBP attending outpatient physical therapy.

Our study was a reliability study, and the values for minimum detectable change represent estimates of measurement error. These values are not to be confused with the minimal clinically important difference (MCID).[26] *Minimal clinically important difference* has been defined as "the smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management."[26(p408)] Clinical decision making is impeded when minimum detectable change exceeds the

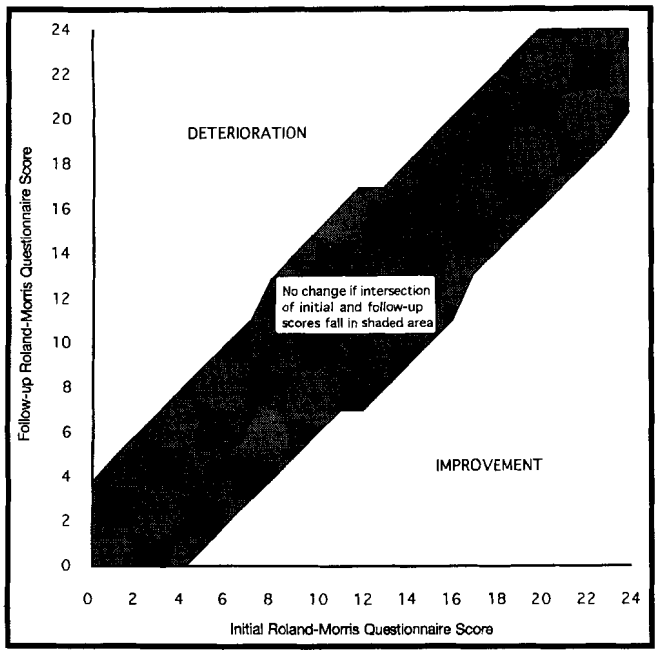

Figure.
Illustration of minimum detectable change at the 90% confidence level.

MCID. There are currently no reports that identify the MCID for the RMQ or for any of the disability measures applied to patients with LBP. We believe that future inquiry should attempt to estimate the MCID and determine the extent to which it is dependent on patients' initial scores.

Conclusion

This work calculated CSEM to estimate minimum levels of detectable change in RMQ points for patients with LBP. The magnitude of minimum detectable change, 4 to 5 RMQ points determined at the 90% confidence level, is dependent on the scores being compared. The results of our study indicate that improvement in patients with initial scores lower than 4 RMQ points and deterioration in patients with initial scores greater than 20 RMQ points cannot be detected with a high degree of confidence. Ongoing challenges include defining the MCID and identifying strategies for detecting improvement in patients with low levels of disability and deterioration in patients with high levels of disability.

References

1 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.

2 Roland M, Morris R. A study of the natural history of back pain, part II: development of guidelines for trials of treatment in primary care. *Spine.* 1983;8:145–150.

3 Fairbank JC, Couper J, Davies JB, O'Brien JP. The Oswestry low back pain disability questionnaire. *Physiotherapy.* 1980;66:271–273.

4 Waddell G, Main CJ. Assessment of severity in low-back disorders. *Spine.* 1984;9:204–208.

5 Ware JE, Sherbourne CD. The MOS 36-item short-form health survey (SF-36), 1: conceptual framework and item selection. *Med Care.* 1993; 31:247–263.

6 Kopec JA, Esdaile JM, Abrahamowicz M, et al. The Quebec back pain disability scale: measurement properties. *Spine.* 1995;20:341–352.

7 Järvikoski A, Mellin G, Estlandre AM, et al. Outcome of two multimodal back treatment programs with and without intensive physical training. *J Spinal Disord.* 1993;6:93–98.

8 Hsieh CJ, Phillips RB, Adams AH, et al. Functional outcomes of low back pain: comparison of four treatment groups in a randomized controlled trial. *J Manipulative Physiol Ther.* 1992;15:4–9.

9 Stratford PW, Finch E, Solomon P, et al. Using the Roland-Morris Questionnaire to make decisions about individual patients. *Physiotherapy Canada.* In press.

10 Deyo RA. Comparative validity of the sickness impact profile and shorter scales for functional assessment in low-back pain. *Spine.* 1986; 11:951–954.

11 Grönbald M, Jupli M, Wennerstrand P, et al. Intercorrelation and test-retest reliability of the pain disability index (PDI) and the Oswestry disability questionnaire (ODQ) and their correlation with pain intensity in low back pain patients. *Clin J Pain.* 1993;9:189–195.

12 Ruta DA, Garratt AM, Russell IT. Developing a valid and reliable measure of health outcome for patients with low back pain. *Spine.* 1994;19:1887–1896.

13 Ruta DA, Garratt AM, Leng M, et al. A new approach to the measurement of quality of life: the patient-generated index. *Med Care.* 1994;32:1109–1126.

14 Stratford PW, Binkley J, Solomon P, et al. Assessing valid change over time in patients with low back pain. *Phys Ther.* 1994;74:528–533.

15 Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. *J Chronic Dis.* 1986;11:897–906.

16 Bergner M, Bobbitt RA, Carter WB, et al. The Sickness Impact Profile: development and final revision of a health status measure. *Med Care.* 1981;19:787–805.

17 Ottenbacher KJ, Johnson MB, Hojem M. The significance of clinical change and clinical change of significance: issues and methods. *Am J Occup Ther.* 1988;42:156–163.

18 Weber H, Holme I, Amlie E. The natural course of acute sciatica with nerve root symptoms in a double-blind placebo-controlled trial evaluating the effect of piroxicam. *Spine.* 1993;18:1433–1438.

19 Herman E, Williams R, Stratford PW, et al. A randomized controlled trial of transcutaneous electrical nerve stimulation (Codetron) to determine its benefits in a rehabilitation program for acute occupational low back pain. *Spine.* 1994;19:561–568.

20 Feldt LS, Brennan RL. Reliability. In: Linn RL, ed. *Educational Measurement.* New York, NY: Macmillan Publishing Co; 1989:108–127.

21 Lord FM. Standard errors of measurement at different score levels. *Journal of Educational Measurement.* 1984;21:239–243.

22 Keats JA. Estimation of error variances of test scores. *Psychometrika.* 1962;27:59–72.

23 Streiner DL, Norman GR. *Health Measurement Scales: A Practical Guide to Their Development and Use.* Oxford, England: Oxford University Press; 1989.

24 Bowman SJ, Wedderburn L, Whaley A, et al. Outcome assessment after epidural corticosteroid injection for low back pain and sciatica. *Spine.* 1993;10:1345–1350.

25 Hadler NM, Curtis P, Gillings DB, et al. A benefit of spinal manipulation as adjunctive therapy for acute low-back pain: a stratified controlled trial. *Spine.* 1987;12:703–706.

26 Jaeschke R, Singer J, Guyatt GH. Measurement of health status ascertaining the minimal clinically important difference. *Control Clin Trials.* 1989;10:407–415.

Appendix.

Sample Calculation of the Conditional Standard Error of Measurement (CSEM)

Step 1. Determine the sample mean and standard deviation for initial and follow-up visits.
Initial visit: 11.5±6.3 Follow-up visit: 6.57±5.8

Step 2. Kuder-Richardson 20 (KR_{20}) reliability coefficient for initial and follow-up visits.
Initial visit: .92 Follow-up visit: .92

Step 3. Kuder-Richardson 21 (KR_{21}) reliability coefficient for initial and follow-up visits.
Initial visit: .89 Follow-up visit: .90

Step 4. Calculate conditional error variances and CSEMs for all Roland-Morris Questionnaire (RMQ) scores using the correction factor suggested by Keats.[22]

An example for a score of 14 RMQ points is

$$\sigma^2_{E_{14}} = \left[\frac{(n-X_p)\,(X_p)}{n-1}\right]\left[\frac{1-KR_{20}}{1-KR_{21}}\right]$$

where $\left[\frac{(n-X_p)(X_p)}{n-1}\right]$ represents the error variance based on the binomial theory of measurement error, $\left[\frac{1-KR_{20}}{1-KR_{21}}\right]$ specifies the correction factor suggested by Keats,[22] $\sigma^2_{E_{14}}$ represents the error variance for an RMQ score of 14, $X_p = 14$, and n equals the number of RMQ items (24).

$$\sigma^2_{E_{14}} = \left[\frac{(24-14)(14)}{24-1}\right]\left[\frac{1-0.915}{1-0.886}\right]$$

$$\sigma^2_{E_{14}} = 4.54$$

$$CSEM_{14} = \sqrt{\sigma^2_{E_{14}}}$$

$$CSEM_{14} = 2.13 \text{ RMQ points}$$

Step 5. Error variances for RMQ scores of 0 and 24 were estimated by dividing the minimum error variance for RMQ scores between 1 and 23 by 2.

Step 6. Determine minimum level of detectable change at the desired confidence level of interest. For exomple, if one wishes to determine whether a person who had an initial score of 14 and a follow-up score of 6 represented a true change at the 95% confidence level, the following method is used:

$$Z = \frac{\text{Initial RMQ Score} - \text{Follow-up RMQ Score}}{\sqrt{\sigma^2_{E_{14}} + \sigma^2_{E_6}}}$$

$$Z = \frac{14-6}{\sqrt{4.54+3.88}}$$

$$Z = 2.76$$

The Z value associated with the 95% confidence level is 1.96. Given that 2.76 is greater than 1.96, it can be concluded that the patient has undergone a true change.

Invited Commentary

My comments on the article of Stratford and colleagues deal with two major issues. The first issue relates to the statistical approach used by the authors to describe the error associated with Roland-Morris Questionnaire (RMQ) change scores. The second and most important issue relates to the application of the current report and related work of Stratford and colleagues to clinical practice.

When attempting to document whether a patient's level of disability (or any other attribute) has changed, the therapist has to be concerned about the error present in both the initial and follow-up measurements. The initial and follow-up measurements are compared to derive a change score. It is this change score that is important for clinical decision making. Many of our clinical decisions are based on comparisons of measurements of an attribute taken during a patient's care. This report is one of the few in our literature that establishes the magnitude of error associated with change scores.

Stratford and colleagues referenced their earlier work that demonstrated that the standard error of the measurement (SEM) at the 95% confidence level is 5 RMQ points. The SEM is used in the calculation of the Reliability Change Index (RCI).[1] The RCI is a statistical procedure for estimating the error associated with change scores. This earlier work is closely related to the current report, but unfortunately, the earlier work was not yet published at the time this commentary was written. The authors reported that they used the RCI as defined by Ottenbacher et al.[2] Ottenbacher and colleagues, however, also described a revised Reliability Change Index (RCI') that requires the use of the standard error of the difference in the calculation of the index. Ottenbacher et al argued that the standard error of the difference is thought to be more indicative of the error present in change scores because it takes into account the error in both the initial and follow-up scores. The RCI accounts for error in only one of the two measurements used to determine the change score. It is not clear whether the RCI or the RCI' was used by Stratford and colleagues in their earlier work.

I used the raw data reported by Roland and Morris[3] to calculate the standard error of the difference at the 95% confidence level for repeated measurements of the RMQ on 20 patients with low back pain. I found the standard error of the difference at the 95% confidence level to be rounded off to 5 points. Estimates of the error associated with change scores are very similar, not only for the current work and the earlier work of Stratford et al but also for the original data reported by Roland and Morris.

In the current article, Stratford and colleagues also have demonstrated that measurement error varies depending on where the measurements fall on the scale. Error is "conditional" on where the measurement falls on the scale. Measurements that fall nearer to the ends of the scale will generally have less error than measurements that fall near the middle of the scale.[4] Less variability is essentially "built-in" at the ends of the scale; therefore, the error theoretically should be less as compared with measurements that fall near the middle of the scale. As Stratford and colleagues report, the conditional SEM is one method used to account for this variability.[5]

The fact that measurement error varies along a scale is recognized as being important by many groups. The American Psychological Association describes the following in their measurement standards, *The Standards for Educational and Psychological Testing*:

> Standard 2.10. Standard errors of measurement should be reported at critical score levels. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported for score levels at or near the cut score. Comment: Reporting standard errors of measurement at every score level may not be feasible in some circumstances, but they should be reported at appropriate, well-separated levels or intervals.[6(p22)]

The American Physical Therapy Association does not directly address the issue of conditional standard error of measurement, but the Association's Standards for Tests and Measurements in Physical Therapy Practice state the following:

> **S13.5.** Research reports on reliability written by secondary purveyors [researchers] must include a description of the statistics used to derive reliability estimates. The rationale for the use of these statistics must be provided. When methodologically appropriate, reports of confidence intervals and standard errors of measurement should be included. Examples of how the reliability estimates are to be used as part of data interpretation should be included.[7(p28)]

> **U44.2.** Test users must consider the error associated with their measurements when they interpret their test results. Reliability and validity estimates should be considered when the test user makes interpretations of measurements.[7(p42)]

Stratford and colleagues have taken reliability assessment to a much more sophisticated but, paradoxically, a much more user-friendly level. Specifically, the authors have highlighted the need for assessing the reliability of change scores taken on a patient. The authors then used a statistical approach that allows clinicians to simply view a graph (see Figure in their article) to determine when true change in disability has occurred in their patients.

As the authors indicate, however, knowing when true change has occurred is only part of the picture in disability assessment. The clinical importance of the change should be judged.

Perhaps the most critical issue discussed by the authors is the concept of a minimal clinically important difference (MCID).[8] The definition of the MCID put forth by Jaeschke and colleagues and referenced by Stratford et al is the following: "the smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management."[8(p408)] The clinical importance of the MCID seems clear and relates directly to the clinical meaningfulness of the change.[9,10] Roland-Morris Questionnaire scores can change over time, just as any measurement can change. The use of the MCID is an attempt to identify the minimal amount of change that is necessary before concluding that the change was clinically important. The MCID takes us back to the concept of statistical versus clinical significance. Scores may demonstrate a statistically significant change but may not be meaningful clinically. Stratford and colleagues acknowledge the importance of determining how large a change needs to be in order to be judged to be important (have an impact on patient care). The definition of the MCID seems confusing, however, when applied to patients receiving physical therapy. For example, a patient may perceive a certain change in an RMQ score as being beneficial, but the change would not necessarily mandate a modification in the patient's treatment. Is this considered a clinically important change? Could the authors elaborate on their discussion of the MCID and how it might be examined in physical therapy clinical practice? Stratford and colleagues state that clinical decision making is impeded when the minimal detectable change (measurement error) exceeds the MCID. From a measurement standpoint, is it possible for the MCID to be smaller than change attributable to measurement error?

Stratford and colleagues have clearly established through a series of reports that the error associated with RMQ change scores varies between 4 and 5 points, depending on where on the scale the change is occurring. The authors also have alerted us to the notion that detecting improvement in patients with very low levels of disability and detecting a worsening in patients with very high levels of disability is problematic. By using the data of Stratford and colleagues, clinicians can now determine very easily and with confidence when a change in disability, as measured with the RMQ, has occurred. As the authors suggest, identifying when true changes in RMQ scores occur is a necessary but incomplete step in the process of judging the importance of changes in disability. Further study is needed to examine the clinical meaningfulness of measurable changes in RMQ scores.

The article of Stratford and colleagues appears to be the first in a physical therapy publication to address the issue of conditional standard error of measurement when examining reliability. This work and the earlier work of Stratford and colleagues[1] will serve as an excellent model for researchers who hope to further clarify for clinicians how to best account for measurement error in clinical practice.

Daniel L Riddle, PT
Associate Professor
Department of Physical Therapy
Medical College of Virginia Campus
Virginia Commonwealth University
Box 980224
Richmond, VA 23298
(driddle@gems.vcu.edu)

References

1 Stratford PW, Finch E, Solomon P, et al. Using the Roland-Morris Questionnaire to make decisions about individual patients. *Physiotherapy Canada.* In press.

2 Ottenbacher KJ, Johnson MB, Hojem M. The significance of clinical change and clinical change of significance: issue and methods. *Am J Occup Ther.* 1988;42:156–163.

3 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.

4 Feldt LS, Steffen M, Gupta NC. A comparison of five methods for estimating the standard error of measurement at specific score levels. *Applied Psychological Measurement.* 1985;9:351–361.

5 Feldt LS, Brennan RL. Reliability. In: Linn RL, ed. *Educational Measurement.* New York, NY: Macmillan Publishing Co; 1989:108–127.

6 American Educational Research Association, American Psychological Association, National Council on Measurement in Education. *The Standards for Educational and Psychological Testing.* Washington, DC: American Psychological Association; 1985:22.

7 Standards for Tests and Measurements in Physical Therapy Practice. In: Rothstein JM, Echternach JL, eds. *Primer on Measurement: An Introductory Guide to Measurement Issues.* Alexandria, Va: American Physical Therapy Association; 1993. Appendix.

8 Jaeschke R, Singer J, Guyatt GH. Measurement of health status ascertaining the minimal clinically important difference. *Control Clin Trials.* 1989;10:407–415.

9 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.

10 Lydick E, Epstein RS. Interpretation of quality of life changes. *Qual Life Res.* 1993;2:221–226.

Author Response

We would like to thank Mr Riddle for his comments and particularly for the clarity with which he addressed several concepts raised in our article. In addition, we have been asked to elaborate on several issues.

The first issue relates to the method used to calculate the Reliability Change Index (RCI) referred to in an article that is in press.[1] The standard error of measurement at the 95% confidence interval of 5 points on the Roland-Morris Questionnaire (RMQ) referred to in previous work[1] was based on the standard error of the difference score (ie, RCI').[2] It was calculated as $1.96 \times \sqrt{2MSE}$, where 1.96 is the tabled z value representing the 95% confidence interval (two-tailed) and MSE is the mean square error term from a test-retest reliability study. The value of MSE was 3.2, and this value yielded an estimate of 5 RMQ points.

The second issue addresses the concept of a minimal clinically important difference (MCID). The definition of MCID provided by Jaeschke and colleagues[3] was intended to be applied to between-group comparisons rather than within-patient comparisons (GH Guyatt, personal communication, November 1995). The application of the concept of MCID to an individual patient requires further clarification. When faced in clinical practice with a change in a measure, clinicians may pose two questions: (1) Has the patient demonstrated a true change (ie, greater than the minimal detectable change)? and (2) Is the magnitude of the change important to the patient? The answers to these questions contribute to a clinician's determination of a course of action. Issues considered by the clinician in making a judgment include the amount and direction of the change, the probable cause of the change, and whether the magnitude of the change is as expected for the period over which the change was assessed.

The first question can be answered by comparing the observed change in the patient with the magnitude of measurement error. The results presented in our report allow this comparison, and a decision can be made without knowing the magnitude of a clinically important difference. The second question, which addresses the importance of the change to the patient, or the MCID, is more challenging. We believe that the magnitude of the MCID is dependent on a patient's initial disability level. For example, when a disability questionnaire is used to assess patients who have low levels of disability (low scores on the RMQ), a clinically important improvement may occur if only one or two activities improve, whereas patients with high levels of disability may require a greater number of activities to improve in order to achieve a clinically important difference. This phenomenon may occur due to regression toward the mean. One strategy for estimating the MCID would be to administer the RMQ at two points in time. For example, the first measurement would be obtained during the initial assessment, and the patient would be asked to prompt the clinician to administer the questionnaire a second time when the patient felt that a small yet important change had occurred. The difference between the initial and follow-up scores would represent an estimate of the MCID. Because the MCID is likely related to the initial amount of patient disability, multiple estimates of MCID could be determined for subgroups of patients with a number of initial RMQ score ranges (eg, 0–6, 7–12, 13–18, 19–24).

Riddle asked whether it is possible for the MCID to be smaller than change attributable to measurement error. If a strategy similar to that mentioned above were used to define the magnitude of the MCID, it would be possible to obtain a value for the MCID that is smaller than the magnitude of the minimal detectable within-patient standard error.

Defining, quantifying, and disseminating estimates of minimal detectable change and MCID pose many challenges. Obtaining a consensus to the questions posed by Mr Riddle will provide the first step to an interesting and hopefully clinically useful course of inquiry.

Paul W Stratford, MSc, PT
Jill Binkley, MClSc, PT, COMP
Patricia Solomon, PhD, PT
Elspeth Finch, MHSc, PT
Caroline Gill, PT
Julie Moreland, MSc, PT

References

1 Stratford PW, Finch E, Solomon P, et al. Using the Roland-Morris Questionnaire to make decisions about individual patients. *Physiotherapy Canada.* In press.

2 Ottenbacher KJ, Johnson MB, Hojem M. The significance of clinical change and clinical change of signficance: issues and methods. *Am J Occup Ther.* 1988;42:156–163.

3 Jaeschke R, Singer J, Guyatt GH. Measurement of health status ascertaining the minimal clinically important differences. *Control Clin Trials.* 1989;10:407–415.