Answers

Methods to assess responsiveness

1. The authors did not define in advance the expected correlations between the change scores. Based on the (dis)similarities of the constructs, they should have stated the expected magnitude and direction of the correlations. It is difficult to interpret these correlations afterwards, and to decide whether the low correlations are due to differences in the constructs being measured or to lack of responsiveness of the instruments under study. Furthermore, the correlations are also influenced by the reliability and validity of the instruments that are used for comparison. In addition, limited reliability or (construct) validity of the SIP or the RDQ may also explain the low correlations.

2. The SIP-PD has a higher t-statistic than the RDQ. However, this does not mean that the change scores of the SIP-PD are more valid than the change scores on the RDQ. The SIP-PD probably had less measurement error, because it has more items than the RDQ (45 compared to 24). Note that the differences between the instruments are very small, and could also be coincidental.

3. Table 7.6 illustrates the significance of change scores, and Table 7.7 concerns the ability of the instrument to distinguish between patients who were considered to be improved versus not improved. This latter approach focuses on the *validity* of change scores, while the former approach focuses on the *statistical significance* of the change scores.

4. As indicated by the authors, the criteria applied in this study might not be appropriate gold standards. The patients were asked to indicate their ‘pain improvement’ on a 6-point scale. This pain improvement may be a different construct than functional status measured with the SIP questions or disability measured with the RDQ. The clinician was instructed to consider overall improvement, based on the patient’s appearance, self-rating, and physical examination. Again, this may be a different construct than functional status or disability.

5. The design of the study could be improved in four ways: first by administering the SIP-PD and the RDQ as two separate measurement instruments, instead of calculating the RDQ scores from the SIP. Secondly, the design could be improved by changing the formulation of the patient's and the clinician's rating of change. It would be better to ask about change in functional status (change in performing daily activities), instead of improvement in pain, because the SIP and RDQ aim to measure (changes in) functional status. Thirdly, hypotheses could be formulated about expected correlations between change scores and about expected differences in scores between improved and not improved patients. For example, it could be hypothesized that the RDQ would correlated at least 0.10 higher than the SIP-PD with the patient's and the clinician's rating of change. Or, it could be hypothesized that patients who have fully resumed all activities had at least a 10% greater change in score, compared to patients who were unchanged. It could also be hypothesized that changes in RDQ scores would be greater than changes in SIP-PD scores (e.g. a difference of 0.10 in effect size), because the RDQ is disease-specific and therefore expected to be more responsive. And fourthly, more emphasis should be laid on the patient's rating of change as a criterion, than on the clinician's rating of change because the SIP-PD and the RDQ are both patient-reported outcomes. Therefore, the patient's rating of change should be considered as a better criterion than the clinician's rating of change. It could be hypothesized that changes in the SIP-PD and the RDQ would correlate at least 0.10 higher with the patient's rating of change than with the clinician's rating of change.