Qual Life Res (2007) 16:815–822
DOI 10.1007/s11136-007-9180-x

The mathematical relationship among different forms of responsiveness coefficients

G. R. Norman · Kathleen W. Wyrwich ·
Donald L. Patrick

Received: 24 January 2006 / Accepted: 18 January 2007 / Published online: 10 March 2007

Abstract

Background Little consensus exists regarding the most appropriate measure of responsiveness. While most indices are variants on Cohen's effect size, the mathematical relationships among these indices have not been elucidated. Consequently, the health-related quality of life (HRQL) literature contains many publications in which a variety of different indices are computed and differences among them noted. These differences are completely predictable when the underlying analytical form of each coefficient is explicated.

Methods In this paper, we begin with a mathematical analysis of the variance components underlying an observed change score. From this, we determine analytically the relationships among the more commonly used indices of responsiveness.

Conclusions Based on this analysis, we conclude that Cohen's effect size and the Standardized Response Mean are the two most appropriate measures, as each provides unique information and each best captures an important relation between treatment effect and variability in response. However, the latter should be interpreted with caution, as under some circumstances, any measure based on variability in change scores can give misleading information. On this basis, we recommend that future analysis of responsiveness be restricted to the Cohen effect size to ensure interpretability and comparability with treatment effects in other domains.

Keywords Responsiveness · HRQL · Equivalence · Mathematics

In 1985, Kirshner and Guyatt [11] proposed a new index for assessing the usefulness of a health-related quality of life [HRQL] measure, which they called ''responsiveness''. If the goal of therapy is, in most cases, to improve HRQL, (or to maintain HRQL in a situation of a chronic degenerative disease), it was reasonable to examine the extent to which an HRQL measure was sensitive to changes induced by therapy as a useful addition to the standard psychometric indices of reliability and validity. Although some authors have disputed whether responsiveness is really a new criterion or can be simply included as one aspect of reliability [13] or validity [10], most studies of HRQL measures include some reference to responsiveness. Interestingly, the notion of responsiveness, sensitivity to change, or the ability to detect change as a critical component of instrument validation appears to be unique to measurement of HRQL; highly regarded standards such as those of the American Psychological Association [1] give no mention of responsiveness to change as an essential aspect of validity.

Although many within the field of HRQL research would argue that responsiveness is important, little consensus exists regarding how to measure it. Confusion exists at the level of conceptualization, study design and measurement. Conceptually, some authors view responsiveness as simply an index of how much change occurred as a

G. R. Norman (✉)
Department of Clinical Epidemiology and Biostatistics, MDCL 3519, McMaster University, 1200 Main St. W., Hamilton, ON L8N 3Z5, Canada
e-mail: norman@mcmaster.ca

K. W. Wyrwich
Department of Research Methodology, Saint Louis University, St. Louis, MO 63103, USA

D. L. Patrick
Department of Health Services, University of Washington, Seattle, WA 98195-7660, USA

consequence of treatment. Others have incorporated interpretation of the change and framed responsiveness as the ability to detect some minimal change, as defined by some external criterion, called a ''Minimally Important Difference'' or a Minimum Clinically Important Difference''. Differences of definition abound; ''minimal'' can be viewed as a ''just noticeable difference'', i.e. any changed that can be detected by the participant, or as some chnage that is important to patients and clinicians, however importance is defined. In this paper, we will take the more inclusive view that responsiveness is an index of ability to detect any change.

In the area of study design, initial studies of responsiveness examined change induced by treatments of known effectiveness, such as hip replacement or cataract surgery, to determine the extent to which an instrument could detect the change. However, this is problematic on both theoretical and practical grounds. The conceptual problem is that there is no defensible way to standardize the amount of treatment, and any measure will show a larger responsiveness to a very effective treatment than to a less effective treatment. The practical problem is simply that in order to examine responsiveness to an effective treatment, the investigator must conduct a trial of a therapy of known effectiveness, which raises both ethical and logistical issues. One alternative is to simply assess patients longitudinally on two or more occasions, ask patients to retrospectively identify whether they have or have not changed and by how much, then determine the difference between changed and unchanged groups, again using some standardized index. However as Norman et al. [15] pointed out, such an approach confounds differences within groups with treatment effects.

Another area of continuing debate is the appropriate statistic to measure responsiveness. In contrast to reliability, where there has long been consensus that an intraclass correlation (or its equivalent, weighted kappa [23]) is most appropriate, a number of responsiveness statistics have been proposed. In perhaps the most exhaustive review of the area, Terwee et al. [24] identified 31 different measures of responsiveness. They pointed out that different measures lead to different conclusions, while stressing that many of these differences reflect different underlying conceptualizations of responsiveness.

One consequence of this dilemma of differing methodologies to assess responsiveness is that numerous studies compute more than one responsiveness statistic [15] to determine which is larger, but provide no basis for choice of a specific methodology beyond the magnitude of the results. A number of review articles [5–7] have examined various approaches, and again, supply no basis for choice, but advocate the use of multiple methods [24].

Most of the proposed responsiveness coefficients are conceptual variants on an effect size, (mean/standard deviation) [4] and have several common features. However, a number of coefficients, in particular those associated with the definition of responsiveness as ability to detect any change [24], use a statistical test such as a paired *t* test, an unpaired *t* test [3], or equivalently, a significant effect of time in a repeated measures ANOVA [2]. While such coefficients may be useful within a study to compare one instrument to another with the same sample size and patient population, they are not useful as an attribute of the instrument in general, because their magnitude depends on sample size, so must be converted to a sample size free index like an effect size, and we will briefly these conversions later. Similarly, coefficients which rely on the use of a Pearson correlation with changes in other variables as surrogates for ''real change'' are of little use to assess responsiveness of a HRQL measure. Such coefficients, because they depend on the variance in both measures (and hence the heterogeneity of changes in the sample under study) as well as their association, are too study-specific to provide useful information as an index of a measure. Therefore these indices will not be considered further. Parenthetically, as we will discuss later, it is likely simplistic to view any numerical estimate of change under any circumstances as uniquely a property of the instrument; rather it reflects the use of the instrument in a particular situation. In this respect, we parallel discussions of reliability in other literatures [1].

There is one last class of responsiveness measures. Deyo and colleagues [7, 6] have approached the issue of responsiveness in a different manner altogether. Like many of the other methods, he identifies a group who has changed and a second group who has not changed, based on some external measure. He then calculates a Receiver Operating Characteristic (ROC) curve. The ROC approach does not involve any distributional assumptions, so in the general case, cannot be mathematically equated to any of the effect size-based measures. However, if one does assume a normal distribution of scores in the changed and unchanged groups, then the difference between the two groups can be expressed as an effect size, called ***d'***. Thus, in principle, there is a one to one correspondence between the effect size and the parameters of the ROC curve.

Returning to coefficients with the form of an effect size, all involve, in some way or another, a ratio of the change observed in a group over time or a difference between groups, to a measure of variability such as a standard deviation. However, coefficients differ primarily in the choice of the denominator; the standard deviation at baseline, the standard deviation of difference scores, the standard error of measurement, the standard deviation at baseline of a stable group, etc. Not surprisingly, for a given set of data, different coefficients can give substantially different results. However, as we will show, an examination of the mathematical form of each coefficient permits a precise expression of the similarities and differences among the coefficients, the relative magnitude of each, and eventually the assumptions underlying each and the strengths and weaknesses. To begin, we must carefully analyse the components contributing to the error in any computed change score.

Variance components in the measurement of change

Consider an individual patient who is assessed with a HRQL measure before and after a course of treatment. Considering first a measurement prior to treatment, a baseline score, the jth observation on subject **i** (where in this case, **j** = 1), this can be written, using the notation of Classical Test Theory, as:

$$O_{ij} = p_i + e_{ij} \tag{1}$$

whereO_{ij} is the observed score, which consists of a term p_i representing the true score of patient i, and an error term, e_{ij} associated with the jth observation on patient i, i.e. measurement error.

From this equation, it follows that the variance of baseline scores is a sum of variances due to differences between subjects and measurement error (details of this and subsequent equations are in Appendix 1).

$$\sigma^2_{\text{baseline}} = \sigma^2_p + \sigma^2_e \tag{2}$$

where σ^2_p is the variance due to differences between patients and σ^2_e is the variance due to measurement error, which is frequently referred to as the Standard Error of Measurement. The standard deviation of baseline scores then follows directly:

$$SD_{\text{baseline}} = \sqrt{\sigma^2_p + \sigma^2_e} \tag{3}$$

If we consider now a test–retest situation, where baseline scores are subtracted from post-intervention scores, then differences between subjects disappear, but error of measurement arises twice, once from baseline and once from post-intervention. As we show in Appendix 1, the total variance is now:

$$\sigma^2_{\text{pre-post}} = \sigma^2_e + \sigma^2_e = 2\sigma^2_e \tag{4}$$

The standard deviation of the difference scores is then the square root of this expression:

$$SD_{\text{pre-post}} = \sqrt{2\sigma^2_e} = \sqrt{2}\sigma_e \tag{5}$$

There is a simple relationship between test–retest reliability, R, the baseline standard deviation, SD_{baseline} and the Standard Error of Measurement, σ_e, which follows directly from the equation for reliability [22], and which will be relevant to future considerations.

$$\sigma_e = SD_{\text{baseline}}\sqrt{(1-R)} \tag{6}$$

Finally, if we compute a change score after treatment, the error variance resembles the pre-post variance, but now contains an additional term from the ‘‘patient × Treatment’’ interaction.

$$\sigma^2_{\text{change}} = \sigma^2_e + \sigma^2_e + \sigma^2_{p\times T} = 2\sigma^2_e + \sigma^2_{p\times T} \tag{7}$$

and the standard deviation of change scores is:

$$SD_{\text{change}} = \sqrt{2\sigma^2_e + \sigma^2_{p\times T}} \tag{8}$$

The purpose of this paper is to use these expressions to examine the mathematical basis of a number of more common responsiveness coefficients and draw inferences about the relation among the coefficients.

Responsiveness coefficients

In our review, we have identified a number of coefficients that all resemble an effect size in form, but differ in specific choice of the numerator and denominator. The first and most straightforward example is the effect size, defined by Cohen [26] as simply the average change from pretest to post-test divided by the standard deviation at baseline. Other coefficients tend to use different indices of variation in change such as the standard deviation of change scores, or multiply by some constant. As one example, Guyatt's Responsiveness Coefficient [9] uses the standard deviation of change scores of a stable group; as another, the Standardized Response Mean [12] divides by the standard deviation of the change scores. All the measures (with the exception of Deyo's ROC curve, [2])can be grouped into three classes: (1) those that are based on the variability of baseline scores (i.e. differences among patients), (2) those that are based on variability in change scores, and (3) those based on statistical tests.

Measures based on variability in baseline scores

Cohen's effect size

The effect size was defined by Cohen [4] as the difference resulting from treatment (either a simple mean change score or difference between treatment and control group change scores), which simply equals the posttest mean minus the pretest mean, divided by the standard deviation of baseline scores. From Eq. 3, **SD**$_{\text{baseline}}$ contains two components, the variance due to true differences among patients and error variance.

$$ES = \frac{(\overline{X}_{\text{posttest}} - \overline{X}_{\text{pretest}})}{SD_{\text{baseline}}} = \frac{(\overline{X}_{\text{posttest}} - \overline{X}_{\text{pretest}})}{\sqrt{\sigma_p^2 + \sigma_e^2}} \tag{9}$$

The standardized effect size and the normalized ratio

Two variants on this statistic have been proposed, the ''standardized effect size'' (SES) [8, 17], the ''Normalized Ratio'' (NR) [19]. Both are based on a pretest–post-test control group design, but differ in the baseline standard deviation is used.

$$SES = \frac{(\overline{X}_{\text{posttest}} - \overline{X}_{\text{pretest}})}{SD_{\text{baseline(improved)}}} = \frac{(\overline{X}_{\text{posttest}} - \overline{X}_{\text{pretest}})}{\sqrt{\sigma_p^2 + \sigma_e^2}} \tag{10}$$

$$NR = \frac{(\overline{X}_{\text{posttest}} - \overline{X}_{\text{pretest}})}{SD_{\text{baseline(stable)}}} = \frac{(\overline{X}_{\text{posttest}} - \overline{X}_{\text{pretest}})}{\sqrt{\sigma_p^2 + \sigma_e^2}} \tag{11}$$

Under the assumption that patients are randomized to treated and untreated groups, the baseline true and error variances can be assumed to be the same, and these coefficients can be assumed to give, on average, identical results to the effect size.

Measures based on error of measurement

The second class of coefficients use either the error of measurement or variants on the standard deviation of change scores in the denominator. All of these coefficients can be viewed as simply special cases of effect size, as mentioned by Cohen [21, p.48]. Nevertheless all differ in specific form of the denominator.

Effect size based on the standard error of measurement

The simplest coefficient (which we will call the Effect Size based on Standard Error of Measurement (SEM) or ES-SEM), is to divide the change by the SEM, σ_e, so that the formula would now become:

$$ES_{\text{SEM}} = \frac{(\overline{X}_{\text{posttest}} - \overline{X}_{\text{pretest}})}{\sigma_e} \tag{12}$$

Interestingly, to our knowledge, such a coefficient has not been proposed, although Wyrwich's work on the Minimally Important Difference [28] showed that the MID appears to be about equal to the SEM. From Eqs. 6 and 9 above, it follows directly that the SEM coefficient can be related back to Cohen's effect size as:

$$ES_{\text{SEM}} = \frac{ES}{\sqrt{(1-R)}} \tag{13}$$

Since $\sqrt{(1-R)}$ is equal to or less than 1, this coefficient will always be the same as or larger than the Effect Size.

Responsiveness statistic

The responsiveness statistic [3] is presumably intended to reflect the variability in change scores, so includes a factor of √ 2 multiplying the SEM; however, it omits the *p x T* interaction term in the change score:

$$RS = \frac{(\overline{X}_{\text{posttest}} - \overline{X}_{\text{pretest}})}{\sqrt{2}\sigma_e} \tag{14}$$

Because of the factor of √ 2, it will be systematically smaller that the effect size based on the SEM (ES$_{\text{SEM}}$).

Guyatt's responsiveness coefficient

Guyatt's coefficient defined responsiveness as the mean change in the treatment group divided by the standard deviation of change in a stable group. In some applications, this stable group was defined by having patients retrospectively declare whether they had or had not changed, a method that has been criticized in the past [15]. The upper limit of this coefficient is the SD(Change) in the control group, which does not contain the *patient* × *Treatment* interaction.

$$RG = \frac{(\overline{X}_{\text{posttest}} - \overline{X}_{\text{pretest}})}{SD_{\text{Change(Control)}}} = \frac{(\overline{X}_{\text{posttest}} - \overline{X}_{\text{pretest}})}{\sqrt{2}\sigma_e} \tag{15}$$

Since the population error variance in treatment and control groups will, in general, be the same, it is expected that this coefficient will be larger than, or equal to the Responsiveness Statistic above.

Standardized response mean (SRM)

The SRM is defined [9] as:

$$\text{SRM} = \frac{(\overline{X}_{\text{post}} - \overline{X}_{\text{pre}})}{SD_{\text{change}}} = \frac{(\overline{X}_{\text{post}} - \overline{X}_{\text{pre}})}{\sqrt{\sigma_{p\times T}^2 + 2\sigma_e^2}} \tag{16}$$

The SRM denominator includes the error variance multiplied by 2 as does the Responsiveness Statistic.

However, it also includes the $p \times T$ interaction, so is systematically smaller than the RS. The difference between the SRM and ES is the use of the standard deviation of change instead of the standard deviation of baseline scores.

Reliable change index (RCI)

Jacobson [20] used a coefficient related to the SRM but included a factor of 1.96 to ensure that a value of 1.0 would correspond to a statistically significant difference for a sample size of 1. From the above, then, the RCI is just:

$$RCI = \frac{(\overline{X}_{\text{posttest}} - \overline{X}_{\text{pretest}})}{1.96 \times SD_{\text{change}}} = \frac{(\overline{X}_{\text{posttest}} - \overline{X}_{\text{pretest}})}{1.96\sqrt{\sigma^2_{p\times T} + 2\sigma^2_e)}} \quad (17)$$

Since the RCI contains a factor of 1.96 in the denominator, it will always be true that the RCI is the smallest of the coefficients based on change scores.

Coefficients based on statistical tests

As we indicated earlier, some authors assess responsiveness by conducting a statistical analysis and computing a statistical test such as a paired or unpaired *t* test. Since these tests are sample size dependent, they can be used within a study to examine relative responsiveness of different measures, but cannot be used to compare across studies. However, since any *t* test is the ratio of the difference to the standard error of the difference, and the standard error is simply the standard deviation of the relevant difference divided by the square root of the sample size, conversion of a t test result to an effect size is simply a matter of dividing by the square root of the sample size. Thus, an unpaired *t* test examining differences between a treated and control group is simply:

$$t_{\text{unpaired}} = ESx\sqrt{n} \quad (18)$$

Similarly, a paired *t* test, examining the difference between a baseline and a post treatment score, is simply related to the Standardized Response Mean:

$$t_{\text{paired}} = SRMx\sqrt{n} \quad (19)$$

Others [25] have used F tests derived from analysis of variance as an expression of responsiveness within a study. While the mathematics is more complex and is omitted, these too can be related back to an equivalent effect size form.

An example

To show how the various coefficients emerge, consider a simple measure of health status, which has been administered three times to a sample of 50 patients. The first administration is a baseline measure. A week later, and before any treatment is undertaken, the measure is re-administered. Treatment then begins, and after 6 weeks, the questionnaire is administered a third time. To work through the example, we generated a set of data using a random number generator in Excel to create observations from a standard normal distribution. A distribution of ‘‘patients’’ was created, with variance = 0.6 (SD = 0.77); three error distributions with variance = 0.4 (SD = 0.63), and a (patient × Treatment) distribution with variance = 0.3 (SD= 0.55). The variance estimates are arbitrary, and are chosen simply to provide plausible values for reliability, responsiveness, etc. Pretest and posttest scores were then computed by adding the first and second error components to the ‘‘patient’’ component. Post treatment scores were computed by adding the third error component and $p \times T$ component to the’’patients’’ component and adding a fixed number, 2.0, to represent the true overall treatment effect.

Variance components were then estimated using the following steps:

Step 1

A repeated measures ANOVA on pretest and posttest scores to estimate the patient variance and error variance. The ANOVA table is shown in Table 1.

The *MS(within)* is equal to the error variance = 0.421, and the variance between subjects follows from:

$$\text{MS}(p) = \sigma^2_e + 2.0 \times \sigma^2_p \quad (20)$$

so $\sigma^2_p =$ (1.694–0.421)/2 = 0.63. These estimates are, of course, reasonably close to the population values from which the distributions were generated.[1]

Step 2

A repeated measures ANOVA on the pretreatment and post treatment scores to estimate the (patient × Treatment) interaction. This ANOVA table is shown in Table 2. From this analysis, it turned out that the estimated treatment effect was 1.87, slightly smaller than the population value of 2.0 used to create the distribution.

[1] These equations for variance components can be determined from standard statistical texts and some measurement books, e.g. Ref. [6].

Table 1 ANOVA table for prepost simulated data with 100 subjects

Source	SS	df	MS	*F*	*p*
Patients (p)	83.0	49	1.694		
Time (T)	0.356	1	0.356	0.845	.36
Error (p × T)	20.62	49	0.421		

Table 2 ANOVA table for change (intervention) simulated data with 100 subjects

Source	SS	df	MS	*F*	*p*
Patients (p)	107.75	49	2.199		
Time (T)	177.60	1	177.60	261.5	.000
Error (p × T)	33.27	49	0.679		

Now this time:

$$\text{MS(Error)} = \sigma^2_{\text{e}} + \sigma^2_{\text{p}\times\text{T}} = 0.679 \qquad (21)$$

so, the estimate of $\sigma^2_{p\ \times\ T}$ is (0.679 – 0.421)/2 = 0.258, somewhat smaller than the population estimate of 0.30. And on this occasion, *MS(patient)* can be used to compute σ^2_p as before, which equals:

$$\text{MS(patient)} = \sigma^2_{\text{e}} + 2\sigma^2_{\text{p}} \qquad (22)$$

so, this time:

$$\sigma^2_{\text{p}} = (2.199 - 0.679)/2 = 0.76 \qquad (23)$$

which is slightly larger than the original estimate of 0.60. Averaging the two estimates, the mean of σ^2_p is (0.63 + 0.76)/2 = 0.695.

We then used these estimates to calculate the various measures of responsiveness. As can be seen in Table 3, the coefficients vary considerably in magnitude, although, since the ''design'' was a single group, pretest–post-test design, the three estimates based on baseline variances were the same since all used the same estimate of baseline variance. The differences among the coefficients based on change scores are in the expected order, with the ES_{SRM} the largest and the RCI the smallest. There is no clear trend between the two families of coefficients, since this depends on the relative magnitude of variance at baseline and variance of change.

Discussion—Comparison among the effect size measures

By expressing the coefficients in terms of the underlying variance components, we have placed the relative magnitude of each on a mathematical basis. If σ^2_p is less than σ^2_e (which amounts to a test–retest reliability less than 0.5) then the ES will be larger than all the measures based on standard deviation of difference scores. While this may seem improbable, it is not at all uncommon in studies of HRQL for the two variance components to be quite comparable. For example, in a recent review, measures of minimal difference using the baseline SD were comparable to measures using SD of change [16]. If σ^2_p is greater than σ^2_e, then there may be less predictable difference between the Cohen ES and the change-based measures.

Examining the change based measures, the formulas demonstrate ES_{SEM} will always be the largest, because it uses only the standard error of measurement in the denominator. The Responsiveness Statistic (RS) will be next largest, as it contrasts the average change to twice the standard error of measurement, and omits the *patient* × *Treatment* interaction from the denominator. The Guyatt coefficient (RG), based on a ''stable'' group, will be identical to RS if the variance of the control group is used as the measure of change in a stable group. To the extent that ''stability'' is assessed by retrospective judgment, RG may be larger than RS. The SRM includes both the variance due to change and the *patient* × *Treatment* interaction, hence is smaller than the previous coefficients. The Jacobson RCI will be the smallest, since it multiplies the denominator of the SRM by 1.96.

Which is ''right'' or ''righter''? First, measures based on baseline differences and measures based on change scores yield different information. The former is a comparison of the average observed change to differences among patients, so one could make a statement about how someone in the 50% percentile of treated patients might compare to untreated patients. The latter family is a comparison of the average change to the variability in change, so is analogous to a statistical test, examining the size of an effect relative to its error. If the goal is to compare the change due to treatment against the expected variability in change, like the usual interpretation of an effect size, then the SRM would appear to be the most unbiased, because it represents the ratio of the average difference to the error in this difference. The SEM and RG ignore the *patient* × *Treatment* interaction, which is a necessary contribution to the variance in response to treatment, and the RCI introduces an additional 1.96 into the denominator.

However, the use of some measure of variability in change scores in the denominator may result in problems of interpretation. If, for example, all patients were to change exactly the same amount, regardless of how large or small the actual treatment effect, then the standard deviation of change is 0, the statistical test is infinitely large and all of these the indices of responsiveness are infinitely large [14]. While this hypothetical example is unlikely to occur in

Table 3 Various indices of responsiveness and calculated value based on simulated data

Coefficient	Abbreviation	Formula	Result
Based on baseline differences			
1) Effect size	ES	Δ/ Pooled baseline SD	1.67
2) Standardized effect size	SES	Δ/ Baseline SD (improved)	1.67
3) Normalized ratio	NR	Δ/ Baseline SD (control)	1.67
Based on differences in change			
4) Effect size (SRM)	ES$_{\text{SEM}}$	Δ/ SEM	2.88
5) Responsiveness statistic	RS	Δ/ √ 2 SEM	2.04
6) Guyatt responsiveness	RG	Δ/ SD$_{\text{change}}$(Control) = Δ / √ 2SEM	2.04
7) Standardized response mean	SRM	Δ/ SD$_{\text{change}}$	1.78
8) Reliable change index	RCI	Δ/(1.96 SD$_{\text{change}}$)	0.91
Based on statistical tests			
9) Unpaired *t* test	t_{unpaired}	Δ / Baseline SD x √n	11.80
10) Paired *t* test	T_{paired}	Δ / √ 2 SEM x √n	14.42

Δ = (Mean post treatment score – Mean baseline score)

reality, nevertheless, the implication of using a measure of variability of change in the denominator means that any change, regardless of clinical importance or practical consequence, could result in a very large index of responsiveness. Further, since there is no consensus about which of the many responsiveness measures based on change is the ''rightest'' one, the proliferation can lead to potential confusion, as there can be as much as a factor of nearly 3 (1.96 × √ 2) between the smallest and the largest.

On this basis, it may be more sensible to anchor observed change against variability at baseline, which is likely less vulnerable to extreme values and more readily interpretable. If so, then the most defensible measure of responsiveness is also historically the first, Cohen's effect size. Use of ES also facilitates interpretation since changes in HRQL measures can then be compared to both standards of size, such as Cohen's small (0.2), medium (0.5) and large (0.8) thresholds and to a large body of literature about treatment based on computed effect sizes. A practical reason to accept the Cohen ES as a standard is that, as HRQL measures become adopted as a useful endpoint for clinical trials, there must be no potential for confusion or misinterpretation of a responsiveness coefficient. Since there is no clear consensus for choice among the change-based measures, and no clear advantage of using such measures, the prudent course might be to remain with the accepted standard, the Cohen effect size.

Finally, regardless of the particular coefficient ultimately chosen, it must be remembered that, however tempting to view the computed coefficient as a measure of the responsiveness of the instrument alone, such an extrapolation is logically unjustified. The measure of responsiveness depends on the average amount of change induced, either by a treatment or by contrasts within subgroups, and a measure of variability, either at baseline or in changes. Each can be large or small within a particular study. Thus, the caveat applied to studies of reliability should be reiterated here. After all is said and done, responsiveness is a measure of a particular instrument *applied to a particular situation and population* and cannot be viewed in any absolute sense.

Appendix 1: Derivation of variance components

As described in the paper, from Classical Test Theory, any observed score is considered to have two components, a true score and an error:

$$O_{ij} = p_i + e_{ij} \tag{24}$$

where O_{ij} is the observed score, p_i represents the true score of patient i, and the error term, e_{ij} is associated with the jth observation on patient I. By definition, errors have a mean of 0, and a standard deviation, σ_e^2. From this equation, the variance of baseline scores is a sum of variances due to differences between subjects and measurement error.

$$\sigma^2_{\text{baseline}} = \sigma^2_p + \sigma^2_e \tag{25}$$

If we now consider a pretest–posttest situation, where there is no treatment effect, the difference between observed pretest and posttest, from Eq. 1, is:

$$D_i = O_{i2} - O_{i1} = (p_i - p_i) + (e_{i2} - e_{i1}) = (e_{i2} - e_{i1}) \tag{26}$$

where the pretest corresponds to j = 1 and the posttest to j = 2. From this, the variance of the difference scores is:

$$\sigma^2_{\text{pre}-\text{post}} = \sigma^2_e + \sigma^2_e = 2\sigma^2_e \qquad (27)$$

If there is a treatment between the pretest and the posttest, Appendix Eq. 3 contains an additional term corresponding to the effect of treatment on patient *i*, which we will call t_i. This can be viewed in turn as the sum of the overall treatment effect, *T*, and the difference between the overall treatment and the response of patient, *i*, which amounts to a ($p \times T$) interaction.

$$\text{Change}_i = T + (t_i - T) + (e_{i2} - e_{i1}) \qquad (28)$$

Since the variance of the overall change, T, is zero, the variance of the change score is then:

$$\sigma^2_{\text{change}} = \sigma^2_e + \sigma^2_e + \sigma^2_{p\times T} = 2\sigma^2_e + \sigma^2_{p\times T} \qquad (29)$$

References

1. American Psychological Association. (1999). Standards for educational and psychological tests.
2. Anderson, J. J., Firschein, H. E., & Meenan, R. F. (1989). Sensitivity of a health status measure to short-term clinical changes in arthritis. *Arthritis and Rheumatism, 32*, 844–850.
3. Chren, M. M., Lasek, R. J., Flocke, S. A., & Zyzanski, S. J. (1997). Improved discriminative and evaluative capability of a refined version of Skindex, a quality-of-life instrument for patients with skin diseases. *Archives of Dermatology, 133*, 1433–40.
4. Cohen, J. J. (1988). *Statistical power analysis for the behavioral sciences* (p. 8). Erlbaum: Hillsdale, NJ.
5. Crosby, R. D., Kolotkin, R. L., & Williams, G. R. (2003). Defining clinically meaningful change in health related quality of life. *Journal of Clinical Epidemiology, 56*, 395–407
6. Deyo, R. A., & Centor, R. M. (1986). Assessing the responsiveness of functional scales to clinical change: An analogy to diagnostic test performance. *Journal of Chronic Diseases, 39*, 897–906.
7. Deyo, R. A., Diehr, P., & Patrick, D. L. (1991). Reproducibility and responsiveness of health status measures: Statistics and strategies for evaluation. *Controlled Clinical Trials, 2*, 142S–158S
8. Fitzpatrick, R., Ziebland, S., Jenkinson, C., & Mowat, A. (1992). Importance of sensitivity to change as a criterion for selecting health status instruments. *Quality in Health Care, 1*, 89–93.
9. Guyatt, G. H., Walter, S. D., & Norman, G. R. (1987). Measuring change over time: Assessing the usefulness of an evaluative instrument. *Journal of Chronic Diseases, 40*, 171–178.
10. Hays, R. D., & Hadorn, D. (1992). Responsiveness to change: An aspect of validity, not a separate dimension. *Quality of Life Research, 1*, 73–75
11. Kirshner, B., & Guyatt, G. (1985). A methodological framework for assessing health indices. *Journal of Chronic Diseases, 38*, 27–36.
12. Liang, M. J., Fossel, A. H., & Larson, M. G. (1990). Comparison of five health status instruments for orthopedic evaluation. *Medical Care, 28*, 632–642.
13. Lindebloom, R., Sprangers, M. A. G., & Zwinderman, A. (2005). Responsiveness: A reinvention of the wheel? *Health and Quality of Life Outcomes, 3*, 8.
14. Norman, G. R., Stratford, P., & Regehr, G. (1997). Methodological problems in the retrospective computation of responsiveness to change: The lesson of Cronbach. *Journal of Clinical Epidemiology, 50*, 869–879
15. Norman, G. R., Stratford, P., & Regehr, G. (1997) Methodological problems in the retrospective computation of responsiveness to change: The lesson of Cronbach. *Journal of Clinical Epidemiology, 50*, 869–879.
16. Norman, G. R., Wyrwich, K. W., & Sloan, J. A. (2003). Interpretation of changes in health-related quality of life: The remarkable universality of half a standard deviation. *Medical Care, 41*, 582–592.
17. O'Keeffe, S. T., Lye, M., Donnellan, C., & Carmichael, D. N. (1998). Reproducibility and responsiveness of quality of life assessment and six minute walk test in elderly heart failure patients. *Heart, 80*, 377–382.
18. Pickard, A. S., Johnson, J. A., & Feeny, D. H. (2005). Responsiveness of generic health related quality of life measures in stroke. *Quality of Life Research, 14*, 207–219
19. Reilly, M. C., & Zbrozek, A. S. (1992). Assessing the responsiveness of a quality-of-life instrument and the measurement of symptom severity in essential hypertension. *Pharmacoeconomics 2*, 54–66.
20. Sprangers, M. A. G., Moinpour, C. M., Moynihan, T. J., Patrick, D. L., & Revecki, D. A. (2002). Assessing meaningful change in quality of life over time: A user's guide for clinicians. *Mayo Clinic Proceedings, 77*, 561–571
21. Stratford, P. W., Binkley, J. M., & Riddle, D. L. (1996). Health status measures: strategies and analytical methods for assessing change scores. *Physical Theraphy, 76*, 1109–1123.
22. Streiner, D. L., & Norman, G. R. (1993). *Health measurement scales: A practical guide to their development and use* (p. 142). Oxford University Press
23. Streiner, D. L., & Norman, G. R. (2003). *Health measurement scales: A practical guide to their development and use,* 3rd ed. (p. 141). Oxford University Press
24. Terwee, C. B., Dekker, F. W., Wiersinga, W. M., Prummel, M. F., & Bossuyt, P. M. M. (2003). On assessing responsiveness of health-related quality of life instruments: Guidelines for instrument evaluation. *Quality of Life Research, 12*, 349–362.
25. Ware, J. E., Kemp, J. P., Buchner, D. A., Singer, A. E., Nolop, K. B., & Goss, T. F. (1998). The responsiveness of disease-specific and generic health measures to changes in the severity of asthma among adults. *Quality of Life Research, 7*, 235–244
26. Wells, G., Beaton, D., Shea, B., Boers M., Simon, L., Strand, V., Brooks, P., & Tugwell, P. (2001). Minimal clinically important differences: Review of methods. *The Journal of Rheumatology, 28*, 406–412.
27. Wright, J. G., & Young, N. L. (1997). A comparison of different indices of responsiveness. *Journal of Clinical Epidemiology, 50*, 239–246
28. Wyrwich, K. W., Tierney, W. M., & Wolinsky, F. D. (2002). Using the standard error of measurement to identify important changes on the Asthma Quality of Life Questionnaire. *Quality of Life Research, 11*, 1–7.