International Journal of Integrated Care – Vol. 2, 17 December 2002 – ISSN 1568-4156 – http://www.ijic.org/

Statistical significant change versus relevant or important change in (quasi) experimental design: some conceptual and methodological problems in estimating magnitude of intervention-related change in health services research

Dr. Berrie Middel, *Department of Health Sciences, Sub-Division Care Science, University of Groningen, A. Deusinglaan 1, 9713 AV Groningen, The Netherlands*

Dr. Eric van Sonderen, *Department of Health Sciences and Northern Centre for Healthcare research, University of Groningen, The Netherlands*

Correspondence to: Berrie Middel, Phone: +31 50 363 6504, Fax: +31 50 363, E-mail: B.Middel@med.rug.nl

Abstract

This paper aims to identify problems in estimating and the interpretation of the magnitude of intervention-related change over time or responsiveness assessed with health outcome measures. Responsiveness is a problematic construct and there is no consensus on how to quantify the appropriate index to estimate change over time between baseline and post-test designs. This paper gives an overview of several responsiveness indices. Thresholds for effect size (or responsiveness index) interpretation were introduced some thirty years ago by Cohen who standardised the difference-scores (**d**) with the pooled standard deviation ($\mathbf{d}/SD_{pooled}$). However, many effect sizes (ES) have been introduced since Cohen's original work and in the formula of one of these ES, the mean change scores are standardised with the SD of those change scores ($\mathbf{d}/SD_{change}$). When health outcome questionnaires are used, this effect size is applied on a wide scale and is represented as the Standardized Response Mean (SRM). However, its interpretation is problematic when it is used as an estimate of magnitude of change over time and interpreted with the thresholds, set by Cohen for effect size (ES) which is based on SD_{pooled}. Thus, in the case of using the SRM, application of these well-known cut-off points for pooled standard deviation units namely: 'trivial' ($ES < 0.20$), 'small' ($ES \geq 0.20 < 0.50$), 'moderate' ($ES \geq 0.50 < 0.80$), or large ($ES \geq 0.80$), may lead to over- or underestimation of the magnitude of intervention-related change over time due to the correlation between baseline and outcome assessments.

Consequently, taking Cohen's thresholds for granted for every version of effect size indices as estimates of intervention-related magnitude of change, may lead to over- or underestimation of this magnitude of intervention-related change over time.

For those researchers who use Cohen's thresholds for SRM interpretation, this paper demonstrates a simple method to avoid over- or underestimation.

Keywords

responsiveness, health outcome, sensitivity to change, methodology, effect size, standardized response mean, questionnaires

Health-related functional status: concepts, measurement and psychometric properties

Introduction

Methodological problems in estimating change in outcome with well-known measures of quality of life or health status have become a significant place on the research agenda in clinical evaluation research. However, these methodological problems seem to be relevant in integrated care research in which the integrated approach is compared with standard care practice when quality of life or health status outcome measures are used. Furthermore, improving methods to estimate change may contribute in the development of evidence based practice. This article was written because researchers in the field of health services research seem often unaware of the wide variety of indices that may contribute to the understanding of an intervention's or programme's effect (in addition to its statistical significance) in terms of health-related Quality of Life (HRQL) outcome or health-related functional status (HRFS). In the attempts to improve healthcare delivery with a new approach, researchers may have the need to distinguish between those who improved in terms of 'small', 'moderate' or 'large' before this new approach will become general practice. The problem

of testing differences between the new approach-group and a standard care group goes together with the dilemma that with large samples, trivial differences between these groups may be statistically significant.

There is a growing recognition that assessing an intervention's effect should not only focus at the statistical significance of the differences in health outcome between the experimental care and control group, but should also focus at the relevance or importance of these outcomes. Estimating the magnitude of the difference between change scores in both groups, the difference between mean change scores are expressed in standard deviation units with the effect size index (ES). To compare the magnitude of change Δ_E assessed in the experimental group with change Δ_C assessed in the control group the idea of effect size *between* groups can be turned on its side and applied to measurement instruments to estimate the amount of change over time *within* a group. Change over time indices are also applied to measurement instruments to evaluate them in terms of being sensitive to detect change in before—after studies. In literature on psychometrics or clinimetrics the concept of *responsiveness* was introduced to denote the magnitude of change over time or sensitivity to change over time. However, many responsiveness indicators have been proposed and resulted in numerous effect size indices (ES). Most of the indicators agree on the numerator (the change score between baseline and post-treatment) but there is little agreement on the appropriate denominator. Since a general convention for effect size interpretation is used for almost any ES out of this effect size family, researchers run the risk of overestimation or underestimation of an intervention's effect. This paper gives, on the one hand, an overview (not an exhaustive enumeration) of several responsiveness indices that may be relevant for evaluation research in health care. On the other hand, this paper gives a simple solution for underestimation or overestimation for two widely used ES.

Health services research is heavily dependent on valid health measures e.g. of health-related quality of life (HRQL) or health-related functional status (HRFS). These concepts have become important in the measurement of intervention-outcome and used as comparable outcomes in cost-effectiveness evaluation. However, in evaluation studies quality of life outcomes have turned out to be a 'kaleidoscopic' concept since no consensus exists with regard to the meaning of the concept in either the research community or the clinical community. Furthermore, the operationalization of the concept of (health-related) quality of life is heavily dependent on the disciplinary perspective in outcome assessment. This lack of consensus has given rise to the development of a myriad of measures involving different components whose conceptual dimensions vary [1]. Therefore, instruments labelled as quality of life measures ''may appear as health status, physical functioning, emotional functioning, perceived health status, symptoms, mood, need satisfaction, well being, and, often, several of these at the same time'' [2]. During the last 10 to 15 years, there has been an exponential increase in the development and use of instruments to measure the outcomes of medical interventions from the patient's perspective. A family of more than 150 instruments were identified in 75 studies [3]; in 1996, Spilker et al. catalogued nearly 215 measures in their second edition of ''Quality of Life and Pharmacoeconomics in Clinical Trials'' [4]. Since there is no consensus on the theoretical construct of quality of life [2, 5–8], the universe of domains belonging to this concept (and therefore the ongoing discussion on the selection of items by which it is operationalized), we prefer concepts such as health-related functional status. Functional status reflects the ability to perform the tasks of daily life in physical, emotional and social domains. There is also a growing agreement on the components of these constructs and the validity of their measurement; for example, by validating these self-report measures with evidence-based measures [9–11]. By using the term health-related functional status (HRFS) in this paper, we implicitly assume that a change in health status or functioning is indirectly related to the patient's subjective experience of quality of life.

For health care administrators or other health professionals who feel the need to measure HRFS as an outcome in evaluation of, for example, the effectiveness of hip-replacement by comparing integrated care with standard care [12], it is essential to know that the choice of available health status instruments is related to the methodological debate on the psychometric properties of instruments (in contrast to outcomes such as physiologic measures). Consequently, this choice is also associated with methodological issues relating to the interpretation of outcome in terms of the magnitude of intervention-related change over time in HRFS or the assessment of the magnitude of differences in outcome between experimental (e.g. managed care, transmural care, shared care) and control groups (standard care, usual care).

Because improving the functional status of patients has become a central therapeutic goal of treatment for many diseases, it is important that health administrators, clinicians and researchers develop a common understanding of:

- what HRFS concepts mean;
- which measure is likely to be the most appropriate one in the context of the disease and the evaluation of, for example, an interdisciplinary or integrated approach;
- the methods to assess intervention-related change (responsiveness of outcome measures); and
- the methods by which a valid interpretation of the magnitude of that change in terms of relevance or importance can be achieved.

In the current paper, the methods to assess the effectiveness of an intervention in terms of change over time (*responsiveness*) will be discussed since valid assessment of the magnitude of the patients' improvement, deterioration and of no change seems to become important in detecting stable, improved and deteriorated patients-groups to evaluate direct costs of new interventions in the context of disease management.

The psychometric properties of HRFS outcome measurement tools

When the reliability and validity of health-related functioning measures have been established, these psychometric properties are generally accepted conditions for use of these measures in evaluation research.

However, the appropriateness of the instrument designed to measure change over time in persons is not only determined by its reliability and validity. Measuring change in order to evaluate efficacy of, for example, new care interventions requires the instrument to be sensitive to detecting change when patients improve in physical function after that intervention. Over the last 15 years, this property has become well known through the widely used concept of *responsiveness*. Responsiveness of health status measures has been denoted as one of the 'holy trinity' of necessary psychometric properties of health status instruments: reliability, validity and responsiveness although other researchers classify responsiveness as *longitudinal validity* [13]. To quantify responsiveness, several *effect sizes* are used as estimates of the amount of change detected with an instrument. One of the aims of this paper is to address some methodological issues relating to the assessment of change over time in health-related functional status and the meaning of the magnitude of this change in scores *within* experimental and control groups. Traditionally, the many generations of researchers who have evaluated the efficacy of care-related interventions, base their decisions on the statistical significance of the within-group (intervention-related) change over time or any statistically significant difference in change from repeated measurements between experimental (care) and control groups (with the underlying hypothesis that the experimental group should show a higher mean change in terms of improvement compared to the control group) [12, 14]. In some cases, investigators eager for results are likely to detect a statistically significant (but very small) change in scores related to the intervention, simply due to large sample size. Consequently, even if change which is statistically significant, though trivial in magnitude, is detected, the $p < 0.05$ doctrine unwittingly pushes the question of how meaningful, important, relevant, or substantial the change is into the background. Significance tests support the decision as to whether the change is due to chance fluctuation or can be functionally related to (medical) intervention. The observed statistical significance does not indicate the magnitude of change. In spite of this, some researchers implicitly suggest that smaller p-values represent larger, and thus more 'relevant', effects [15].

Against this background, the objectives of this paper can be formulated in terms of the following topics:

- Responsiveness is a construct that is used with different theoretical definitions and with a wide variety of operationalisations by effect size indices.
- How comparable are different operationalisations of effect sizes (ES) when outcome is interpreted as 'trivial' ($ES < 0.20$), 'small' ($ES \geq 0.20 < 0.50$), 'moderate' ($ES \geq 0.50 < 0.80$), or 'large' ($ES \geq 0.80$) according to the well-known thresholds of Cohen? [16]
- How concordant are the effect sizes, labelled by the researcher as 'trivial', 'small', 'moderate', or as 'large' change in a domain of health-related function with the patient's perception of change in the same domain signified with the same qualitative terms?

Responsiveness, a problematic construct

To give greater meaning to the interpretation of the amount of change in scores on health-related functional status instruments, the concept of *responsiveness* was introduced in publications. For evaluation studies, the usefulness of a HRFS- instrument depends on its ability to detect a change that is clinically meaningful. Clinically meaningful refers to a change that justifies alteration in management of the disease or to a change that indicates the efficacy of an innovative type of intervention in domains of HRFS. Responsive measures discriminate between trivial and

substantial changes within groups and consequently, show the difference in change between those groups. Thus, the term *responsiveness* is used as an indicator of the instrument's sensitivity to change, as well as an indicator of the magnitude of intervention-related change over time. The term responsiveness, however, is a confusing one for the beginner who encounters it in the literature, since papers addressing intervention-related change in terms of HRFS may refer to a varying composite of aspects. As appears from a selection of scientific papers, the term *responsiveness* is used as an operational definition of:

- 'An indicator of the sensitivity of an instrument to detect change over time' [17–22] or even refer to the extent to which a measure is sensitive to *real* change [23];
- 'statistically significant change in an experimental group in which change should be present' [24];
- 'an indicator of the magnitude of treatment-related change' [20–22, 25–35, 35–56];
- 'a measure of clinically relevant change in health' [57, 58], although some investigators prefer the term 'clinically *significant* change' [59, 60].

Qualitative terms such as 'clinically important' need at least a golden standard. However, such a standard is not available for HRFS measures. An substitute that is often used for a golden standard for HRFS is an external criterion. The blinded observation of a health professional can be used as an external criterion for justifying the interpretation in terms of clinically relevant or important change in HRFS. Another external criterion or yardstick for the interpretation of changes in HRFS is the patient's perception of the importance of change after (for example) a specific intervention.

Husted et al. [61] distinguished internal responsiveness from external responsiveness by defining internal responsiveness as the ability of a measure to detect change over time, whereas external responsiveness was defined as the extent to which change in a measure relates to corresponding change in a reference measure [11, 62, 63]. Despite this clarification of the concept of responsiveness by this recently published classification, the assessment of change in HRFS over time in evaluation research is quantified using a variety of approaches. For the sake of clarity, we will therefore in this paper use the concepts in the following meaning:

- *responsiveness*: the psychometric property of a measurement instrument, namely its sensitivity to detect difference between two points in time (change over time) within groups;
- *meaningful or relevant difference*: the amount of change in scores or the magnitude of change within and between groups, according to statistical or other quantitative criteria (e.g. effect size indices);
- *clinically* relevant or *clinically* important change in scores on a health-related functional status measure as the magnitude of change that is linked to an external criterion of relevance.

The purpose of a study and its study design may require different psychometric properties of the outcome measure. Consequently, the measure must either have the property of being able to detect differences *between* subjects at a single point in time (discriminative instruments) i.e. the ability to differentiate between groups 'who have a better HRFS and those who have a worse HRFS' [53, 64, 65]. Other studies may require the instrument's ability to detect change over time *within* subjects (evaluative instruments) [66–68]. Consequently, in randomised clinical trials (RCT) or quasi-experimental designs, HRFS-instruments should have both properties, namely: 1. the ability to reliably estimate change between baseline and post-test within an experimental and a control group, and 2. the ability to estimate the difference in change over time by comparing the average change assessed in e.g. patients receiving standard care and in patients receiving the new care intervention in order to determine intervention-related effect, when it is hypothesised that subjects assigned to the care innovation group are expected to change (on the average) more than those in the control group do.

Responsiveness and the instrument's scope: generic versus specific measures

An important criterion for choosing an instrument in order to detect change in HRFS is its generic or disease-specific scope, which will depend on the objectives of the specific study. Generic health status measures seek a broad perspective that is not specifically related to the restricted scope of the HRFS of a specific disease. Therefore, generic measures allow investigators to compare health status across different diseases and interventions [69]. Generic measures are health-related to the extent that disease, injury, treatment, intervention, or policy [70] influences them. Disease-specific measures focus on the disease being studied, allowing greater sensitivity to intervention-related change compared to generic measures. The responsiveness of a health status instrument is an important issue in the decision to use disease-specific or generic measures of health-related functional state.

For example, for those cases in which therapeutic effects are likely to be modest and undramatic [12, 19, 71], a better sensitivity to change over time of an instrument is a necessary condition. In health services research, hypothesising statistically significant change over time and more substantial change (improvement) in patients assigned to the experimental group of managed, shared or integrated care, effects are not likely to be large or impressive. Using disease-specific outcome measures gives an opportunity to tap more precisely intervention-related improvement in domains of health, which may have been deteriorated due to the disease where generic measures contain items that are not likely to be linked to domains of health status that may change due to the disease or handicap of the patients in the study. Although the question of whether instruments, that are tailored to the disease, are superior to measures of general function in terms of sensitivity to change, has not been settled definitely, a growing number of studies indicate that disease-specific measures seem to be more responsive than generic measures [36, 42, 47, 51, 72–76].

Effect size (ES) as indicator of responsiveness

Mean differences in outcomes between baseline and post-intervention of a test can be standardised to quantify a care intervention's effect in units of standard deviation (SD). Consequently, standardising mean change over time with a standard deviation allows comparison of a particular intervention's different outcomes, independent of the measuring units. The resulting statistical measure is known as *effect size* (ES) index. In many evaluation studies, standardised change over time in HRFS (ES) is used in comparisons of groups who were treated differently. This method of expressing change scores in a so-called effect size index seems to be an appropriate method to estimate the magnitude of change over time in before—after study designs.

The effect size index tells us something very different from the p-value, which indicates the obtained probability of a Type I error in a test of statistical significance. If a p-value is annotated as statistically significant, rejecting the null-hypothesis does not imply that the effect was important in any way nor does a non-significant p-value indicate a trivial result [77–80]. Criticism of statistical hypothesis testing has a long history [81], and even Jacob Cohen [15, 82] ''played a prominent role in the anti-hypothesis-testing charge'' [83]. The adoption of a fixed level of significance may lead to the situation in which two researchers obtain identical intervention effects but obtain different p-values (0.04 and 0.06) due to the effect of (slightly) different sample sizes leading to different decisions. Thus, p-values are confounded by the joint influence of sample size and the effect size [84] and make the rejection of the null-hypothesis not very informative. Another criticism of null hypothesis testing is that it is foolish to ask: 'Are the effects of A and B different?' ''They are always different—in some decimal place–for any A and B'' [85]. Since then, quantitative investigators in medical and social sciences have proposed a variety of supplementary effect size indices, some of which we will clarify. Reporting effect sizes without appropriate statistical tests and associated p values is misleading and potentially dangerous if the number of observations that is required to detect a difference has not been estimated by means of a power analysis. Effect size statistics should be provided to supplement statistical testing (not as a substitute for it), and only when the outcome is sufficiently extreme from what would have been expected on the basis of chance ($p < \alpha$).

It should be noted that during the debate on 'significance testing', several vocal leaders in psychology and education research called for the universal reporting and interpretation of empirically produced effect sizes [86, 87].

There are myriad estimates of effect size out of which the researcher can make a choice [88] and the question arises as to which of the effect size measures 'that could be summoned up for a given problem should a researcher report?' [83, 84] The most elegant solution for this problem would seem to be for authors to include the sufficient statistics so that every reader can compute whichever effect size index they believe is best suited to the situation. Table 1 gives an overview of responsiveness measures in repeated measurement study designs.

Estimation of magnitude of change

Effect size: a problematic statistic

For those researchers who are not conversant with this method of estimating the amount of change over time it is essential to know that in the last decade various critical comments about Cohen's work [16] have been made. These include:

- there is no consensus on the 'theoretical' meaning, or the conceptualisation of the effect size as an outcome variable;

Table 1. Formulas for responsiveness measures for change over time (Within-group standardised mean change)

Paired t statistic	$\frac{\bar{X}1-\bar{X}2}{SE^{x}}$
Effect size (1)	$\frac{\bar{X}1-\bar{X}2}{SD^{xx}_{pooled}}$
Effect size (2)	$\frac{\bar{X}1-\bar{X}2}{SD_{baseline\ scores}}$
Effect size (3)	$\frac{(\bar{X}1-\bar{X}2)_{treated\ subjects}-(\bar{X}1-\bar{X}2)_{controls}}{SD_{pooled\ baseline}}$
Standardised Response Mean (1)	$\frac{\bar{X}1-\bar{X}2}{SD_{change\ scores}}$
Standardised Response Mean (2)	$\frac{\bar{X}1-\bar{X}2_{(improved\ subjects)}}{SD_{change\ scores\ (improved\ subjects)}}$
Standardised Effect size	$\frac{\bar{X}1-\bar{X}2_{(improved\ subjects)}}{SD_{baseline\ (improved\ subjects)}}$
Responsiveness index (1)	$\frac{M.C.I.D^{x\,x\,x}}{SD_{change\ scores\ (stable\ subjects)}}$
Responsiveness index (2)	$\frac{\bar{X}1-\bar{X}2}{SD_{baseline\ (stable\ subjects)}}$
Responsiveness index (3)	$\frac{\bar{X}1-\bar{X}2}{SD_{change\ scores\ (stable\ subjects)}}$
Responsiveness coefficient	$\frac{\sigma^2(\bar{X}1-\bar{X}2)}{\sigma^2(\bar{X}1-\bar{X}2)+\sigma^2 error}$
Normalized ratio	$\frac{\bar{X}1-\bar{X}2_{(improved\ subjects)}}{SD_{baseline\ (stable\ subjects)}}$
Relative efficiency statistic	$(\text{t-statistic}_{measure\ 1}/\text{t-statistic}_{measure\ 2})^2$
Relative efficacy index[xxxx]	$(ES_p/ES_{p\ best})^2 \times 100$

[x] SE = standard error of the difference

[xx] where pooled $SD=\sqrt{\frac{(SD_{baseline})^2-(SD_{outcome})^2}{2}}$ for: $N_{baseline}=N_{outcome}$

[xxx] Minimal Clinically Important Difference according to external criterion (i.e. the difference in change score between those who perceived no change and those who perceived little change) which is considered to be the minimal difference in change over time that patient's perceive as meaningful.

[xxxx] Magnitude of change over time is estimated for each scale by dividing the mean change by the pooled variance of change, according to Cohen {154} denoted as ES_p. This relative efficacy statistic is computed by squaring the ratio obtained by dividing each scale ES_p (numerator) by the scale having the largest ES_p (denominator). This statistic is then expressed as a percentage with respect to the best measure.

- there is no consensus on the mathematical way to determine the magnitude of the difference between scores gained on two different occasions: researchers classify the extent of responsiveness and magnitude with effect sizes using several standard deviations (see Table 1);
- threshold values for 'trivial' (<0.20), 'small' ($\geq 0.20 < 0.50$), 'moderate' ($\geq 50 < 0.80$) and 'large' (≥ 0.80) effects only apply to effect size 1 in Table 1.

How to give meaning to the magnitude of change

Regarding the use of the notion of effect size in HRFS research, several researchers have claimed that without an external criterion, the estimated amount of change measured by the effect size index can be denoted as *clinically* important change [20, 21, 57, 58, 89]. Other researchers assume that an effect size, estimated within a group of subjects, expresses the measure's ability to detect change over time (due to an experimental intervention) [17–22, 57, 72] without claiming that their effect size indicates that the instrument is sensitive or responsive to *clinically relevant* changes in the patients' perceived health. When a HRFS instrument is used as an outcome measure, and the amount of change estimated with change scores (or quantified by an effect size) is defined as clinically relevant, the following question logically arises: 'What is meant by a clinically relevant change?' [90, 91] Because patients and health professionals differ in the preferences or perceived relevance that they assign to particular aspects belonging to domains of health-related functional status, several authors have incorporated these perceptions or preferences into health status instruments' items and scales [5, 75, 89, 90, 92–95] to give more significance to the term 'relevant'. In this paper, we address the methodological problems of quantifying change over time

with effect size indices and the risks of overestimation and underestimation according to widely used thresholds introduced some 30 years ago [16].

How to quantify change in terms of effect size

Many evaluation studies have been conducted that use different methods to estimate magnitude of change over time in terms of Effect Size (ES). These have indicated that there is no convincing evidence that either method offers any apparent advantages [6, 74]. The literature shows that numerous quantitative indices belonging to the family of effect sizes (ES) [88] have been developed. However, there is no consensus on how to declare a difference in terms of standard deviation units. The interpretation of the effect size is determined by the choice of the standard deviation used to standardise the mean change over time and, related to that, by the ready adoption of the interpretation guideline as set by Cohen [16]. Several effect size indices are used in HFRS and quality of life research, which have in common that $\bar{X}_1 - \bar{X}_2$ is divided by a standard deviation. The researcher's decision as to which SD he will take is either a well-considered choice or one which is copied from well-reputed colleagues and has no further justification. However, in giving meaning to standardised mean change in terms of 'trivial', 'small', 'moderate', or 'large' effects using the thresholds that Cohen [16] provided us with some thirty years ago, it seems to have been forgotten that these cut-off points were calculated with the *pooled standard deviation* (SD_P). Consequently, applying these thresholds for mean change scores standardised with the standard deviation of the change scores ($\bar{X}_{t-1} - \bar{X}_{t-2}/SD_{X1-X2}$), which is not equal to the pooled standard deviation (SD_P), may lead to over- or underestimates of effects.

For his effect size (mean baseline scores minus mean follow-up scores, divided by the pooled standard deviation) Cohen came up with conventions for those values that constitute a 'trivial' ($ES < 0.20$), 'small' ($ES \geq 0.20 < 0.50$), 'medium' ($ES \geq 0.50 < 0.80$), and a 'large' effect ($ES \geq 80$). However, for each of the effect size and responsiveness indices from Table 1 (except: T-Test, Normalized ratio, and relative- and efficacy indices), these thresholds are used indiscriminately, which may have contributed to the confusion in this area [61].

Effect size interpretation: the threat of internal and external validity of (quasi) experimental research by overestimation or underestimation

In the practice of health-related quality of life research, most researchers remain primarily interested in the statistical significance of the change in health-related functional status or quality of life in pre post designs. In combination with e.g. the T-test approach, substantial effects can be detected [96–98] with an estimate of effect size. If a p-value is annotated as statistically significant, rejecting the null hypothesis does not imply an effect of important magnitude; likewise, a non-significant p-value does not indicate a trivial result [77–80], although some researchers implicitly deem more important those results with smaller p-values.

In the last decade, however, a growing number of longitudinal intervention studies are focussed on questions like ''If the change between baseline and outcome is statistically significant, what can we say about the magnitude (or amount) of change over time that has been detected? Can we interpret this difference in terms of an important difference or as a relevant (substantial) change?'' To answer these questions, the responsiveness i.e. the ability of quality of life outcome measures to detect change over time, has become crucial in the past decade. However, the responsiveness estimation is neglected in many evaluation studies in which it could give information on the importance of change due to intervention-related effects supplementary to the statistical significance of change over time (e.g. before and after intervention) [99, 100]. Reporting effect sizes without appropriate statistical tests and associated p-values is misleading and potentially dangerous when the number of observations that is required to detect a difference has not been estimated with a power analysis. Effect size statistic should be provided to supplement (not as a substitute for) statistical testing, and only then, when the outcome is sufficiently extreme from what would have been expected on the basis of chance ($p < \alpha$).

Noteworthy in this respect is that in the field of psychological research, editorial policy indicates that ''until there is a real impediment to doing so, authors should routinely present an effect size estimate along with the outcome of a significance test'' [84, 86, 87].

Table 1 shows that several quantitative indices have been developed that belong to the family of effect sizes (standardized differences) each calculated with a different denominator in the ($\bar{X}1 - \bar{X}2/SD$) formula, for example, the SD of stable subjects, the SD of the baseline assessment, the SD of the observed change score (improved, stable subjects) etc. Obviously, there is no consensus on how to declare a difference in terms of standard deviation units. Only in a small number of publications is this lack of consensus on the most appropriate effect size indicator signalled [13, 90, 101–104].

Despite the fact that different opinions exist on the method to estimate magnitude of difference between

groups or the magnitude of change within groups, researchers use the straitjacket of thresholds Cohen provided us with some 30 years ago [16]. However, these thresholds are taken for granted by many researchers for every version of effect size index. With regard to the correct use and interpretation of effect size indices as estimates of intervention-related magnitude of change, we must revisit some basic assumptions:

- the ES is developed and elaborated by Cohen to estimate power or the necessary sample size to detect relevant change with the basic principle of independent, equal size samples with common within-population standard deviation $^{\sigma}$;
- in the case that this ES is used to calculate the sample size needed to detect change in paired samples or in a repeated measurement-design it must be adjusted for correct use of Cohen's power tables and sample size tables. However, this adjusted ES cannot be interpreted with Cohen's thresholds for effect size interpretation in evaluation research;

Effect Size estimation with independent (treatment vs. control) and dependent observations (repeated measurement)

Independent samples

Cohen represented the effect size (ES) on some dependent or outcome measure used in an experiment in terms of the difference (using the symbol ***d′*** to denote this ES) between the treatment and control group expressed in units of common within-population standard deviation (in samples this standard deviation is estimated with the pooled standard deviation) as follows:

[Formula A]

$$ES = d' \frac{(\bar{X}_{treatment} - \bar{X}_{control})}{\sigma}$$

With this estimate of effect size, after analysing a wide sampling of behavioural research, Cohen developed his rules of thumb and reported that effect of 0.8^{σ} being on the large end of the range, 0.5^{σ} was the medium, and 0.2^{σ} was at the small end of the range [105].

Dependent samples or paired observations

The difference or change in matched observations within subjects is standardised by the common within-population $^{\sigma}$, according to Cohen's [16] p. 13, but due to the removal of the variation in many extraneous characteristics of the subjects, the index must be adjusted [16], dividing ***d′*** by $\sqrt{(1-r)}$. Cohen used the symbol ***d*** to denote this adjusted ES (in evaluation research often labelled as **S**tandardized **R**esponse **M**ean).

[Formula B]

$$d = \frac{d'}{\sqrt{(1-r)}}$$

d' = effect size for independent samples

d = adjusted effect size

r = correlation between baseline and outcome

This $\sqrt{(1-r)}$ – correction of the denominator of formula A is necessary for a proper use of power and sample size tables since these assume $2(n-1)$ degrees of freedom where, in the case of paired observations, only $n-1$ are actually available [16]. This consequence for power and sample size estimation is something different from the use of the effect size d in evaluating efficacy of a new intervention in terms of amount of change in health status, which was not the aim of Cohen's work.

Overestimation or underestimation of effect by using Cohen's thresholds for SRM

When effect sizes are calculated as the standardized difference in mean score to evaluate the magnitude of difference in HFRS, for example, between an intervention group (interdisciplinary or integrated care and a control group, formula [A] should be used. The effect size can be calculated by pooling the estimates (pooled standard deviation) derived from sample data. In contrast to this independent sample case, effect sizes are also used in evaluation studies (pre- post study designs) as estimates of the responsiveness or change over time within groups. Effect sizes are also in these study designs used to give meaning to change over time in terms of 'trivial' ($ES < 0.20$), 'small' ($ES \geq 0.20 < 0.50$), 'moderate' ($ES \geq 0.50 < 0.80$) or 'large' ($ES \geq 0.80$) change. Cohen [16] introduced this 'matched pairs' effect size, which was later renamed the standardised response mean (SRM) by Liang et al. [106] to avoid confusion concerning other effect size indices. However, several researchers seem to have adopted the idea that **every** standardised difference is subject to Cohen's definitions of trivial, small, moderate and large effect. Such a belief could lead to misinterpretations in studies focussing on intervention-related outcome in paired

Table 2. The conversion of an effect size calculated with the pooled SD (ESP) of 0.42 into an SRM with correlation coefficients ranging from 0.00–0.90

Corr.	0.00	00.10	0.20	0.30	0.40	0.50	0.60	0.65	0.70	0.80	0.86	0.90
$\frac{(0.42/\sqrt{2})}{\sqrt{(1-r)}}$	0.30	0.31	0.33	0.36	0.38	0.42	0.47	0.50	0.54	0.66	0.79	0.94

samples since these cut-off points of the magnitude of the difference were not established as a rule of thumb with the effect size ***d*** (dependent samples) but with the index ***d′*** (independent samples). Thus, we argue that Cohen's thresholds are based on the assumption of common within-standard deviation (with matched pairs sample data we use the raw within-group pooled SD), resulting in an effect size we annotate as ES_P. Consequently, in matched pairs studies these thresholds cannot be used interchangeably for the SRM due to the role of the correlation between repeated measures or between scores from paired samples. In this part of the article the attention is focussed on the standardized change in mean score between two points in time **within** a single group, estimated with the within-group effect size. In relation to the use of Cohen's rule of thumb for effect size interpretation, we evaluate the consequences of the calibration of the SRM with the ES_P and the role of the correlation between pre- and post-test scores.

To investigate how serious discrepancies can appear in effect size interpretation we first elaborate a theoretical example and used a sample of studies to evaluate the seriousness of these differences in practice. To evaluate the seriousness of the discrepancies between SRM and ES_P, the correlation of the subject's repeated measurements was needed. Empirical data were collected for the purpose of secondary analysis to draw conclusions in terms of the relative size of the SRM to the ES_P in relation to the size of the correlation. Applying Cohen's thresholds (which are based on the pooled estimate of effect) to interpret the SRM on the one hand may lead to similar results or subtle and trivial differences, but on the other hand also to meaningful shifts in classification of the amount of estimated change. In this article we analysed 148 SRMs interpreted using Cohen's rule of thumb and compared these SRMs with Cohen's ES_P calculated with the same data. Furthermore, we calculated for the range of the correlation coefficient (***r***) 0.01 to 0.99 the SRM adjusted for Cohen's cut-off points 0.20, 0.50 and 0.80 of the pooled effect size.

To study the consequences of the impact of the association or correlation between repeated measures, we restrict the analysis to two effect size indices suitable for the evaluation and interpretation of magnitude of change over time (or responsiveness) within one group, namely the SRM and the ES_P.

[Formula C]

$$ES_p = \frac{\bar{X}_{change}}{SD_{(pooled)}}$$

The ES_P introduced by Cohen was made comparable to the SRM where the ($SD_{X\text{-change score}}$), is used as the denominator in which, as we will demonstrate below, the correlation between baseline and outcome scores is involved.

The SRM is the ratio between the mean change score and the variability (the standard deviation) of that change score within the same group.

[Formula D]

$$SRM = \frac{\bar{X}_{change}}{SD_{(Xchange-score)}}$$

One of our purposes was to get an indication of how the SRM varies in accordance with the size of the correlation between pre- and post-test scores when the correct pooled effect size estimate is used. An example may illustrate the role of **r**, the correlation of a person's health status measurements over time:

In a study in which the outcome of an intervention was evaluated with a HRFS measure, and in the case of improvement, a lower mean score after intervention was hypothesised. The investigator finds at baseline a mean score of 11.12 with a standard deviation of 4.43 and a mean score of 9.16 (SD: 4.88) at follow up. The estimate of the common within-standard deviation, which is the square root of $(SD_{baseline})^2 + (SD_{outcome})^2/2)$, thus 4.66, and the pooled effect size ***d′*** (ES_P) is then calculated as follows 0.42 (11.12–9.16/4.66). Before we compare the ES_P and SRM in relation to the correlation between repeated measurements, we must solve the problem of the equation of both formulas C and D. According to Cohen, the difference between means for **dependent** samples is standardised by a value "which is $\sqrt{2}\,(1-r)$ as large as would be the case were they independent" [16].

From equation A4 in the appendix $(\mathbf{d}'/\sqrt{2})/\sqrt{(1-r)}$ is equivalent to the SRM and alternatively SRM * $\sqrt{}$ 2 *

Table 3. Comparison of four Standardised Response Means calibrated into Cohen's pooled effect size index (ES_p). Effect Size **d'** (ES_p)

	Trivial	Small	Moderate	Large
SRM	0–<0.20	≥0.20–<0.50	≥0.50–<0.80	≥0.80
Trivial	2			
Small	3	4		
Moderate		9	8	
Large			8	6

$\sqrt{}(1-r)$ is equivalent to ***d'*** and both indices will vary with the size of ***r***. In Table 2, we have elaborated the hypothetical example in which this effect size ***d'*** (ES_P) = 0.42, is transformed into the SRM for a series of values of ***r***. Both effect sizes are equal in the case that **r** = 0.50): $ES_P = (0.42/\sqrt{2})/\sqrt{(1-0.50)} =$ SRM, and the SRM for ***r*** = 0.50 is then $(0.42/1.41)/0.71 = 0.42$. In Table 2 it is shown that the SRM gets larger for larger values of ***r***. For example, an effect size of 0.42 indicating 'small effect' corresponds with a 'medium effect' (SRM = 0.50) if the correlation between the repeated measurements is approximately 0.64. This small effect estimated with the ES_P corresponds with a 'large effect' (SRM ≥ 0.80) if this correlation is approximately 0.86.

If we take Cohen's original work [16] as being valid, we will have to rectify interpretations of the meaning of the estimated magnitude according to the results from this analyses. In previous work, we published two studies [55, 71] in which 40 Standardised Response Mean indices were interpreted according to Cohen's thresholds for pooled estimates of standard deviation (ES_p) out of which 20 turned out to be overestimations or underestimations of intervention-related effect (Table 3).

In another study [107], we analysed this problem using results from other researchers. This secondary analysis of data from other studies revealed that 23% of the estimated effect sizes did not fall in the same magnitude of change category according to the Cohen's thresholds (Table 4).

To avoid invalid interpretations in the evaluation of responsiveness with SRM index we have, for every value of the correlation between baseline and follow-up score, calculated the corresponding ES_P's for Cohen's thresholds of 0.20 = small, 0.50 = medium, and 0.80 = large. Indices that lie within the interval that corresponds with these thresholds are not depicted. To classify the magnitude of change estimated with the SRM more precisely, this effect size index is adjusted for every value of the correlation coefficient (***r***) between baseline and follow-up assessments and brought into line with Cohen's thresholds for effect size. Figure 1 shows that SRMs of 0.20, 0.50 and 0.80, don't deviate after calibration with Cohen's ES_P taken as the original standard, when ***r*** = 0.50. However, an SRM of 0.20 must be tagged as trivial effect as long as the correlation coefficient ranges from ***r*** = 0.01 to ***r*** = 0.49. With large corresponding correlation coefficients (**r** = 0.92) a small SRM of 0.20 must be tagged as moderate $(0.20/\sqrt{2}/\sqrt{1-0.92} = 0.50)$ or (**r** = 0.97) large $(0.20/\sqrt{2}/\sqrt{1-0.97} = 0.80)$. The class midpoint 0.35 of the 'small effect' range of effect (not depicted) has to be classified as moderate or large effect with correlation coefficients of 0.76 $(0.35/\sqrt{2}/\sqrt{1-0.76} = 0.50)$ and 0.91 $(0.35/\sqrt{2}/\sqrt{1-0.91} = 0.80)$, respectively.

SRMs of 0.80 has to be tagged as 'moderate' effect (ES = 0.58–0.79) if the correlation ranges from ***r*** = 0.01 to 0.49. The SRM ≥ 0.80 cannot drop below the cut-off points of small and trivial ES due to the correlation magnitude between baseline and outcome measurements. 'Moderate' effect (SRM = 0.50) must be tagged as 'small' if the correlation between repeated measures is below 0.49 and has to be classified as 'large' (ES ≥ 0.80) in case of ***r*** = 0.81. The class midpoint 0.65 (not depicted) of the 'moderate effect'

Table 4. Similarities and differences between the Standardised Response Mean (SRM) and pooled effect size **d'** (ES_P) interpreted using Cohen's thresholds (n = 148)

ESpooled	ES < 0.20 Trivial effect	ES ≥ 0.20 < 0.50 Small effect	ES ≥ 0.50 < 0.80 Medium effect	ES ≥ 0.80 Large effect	Total
SRM					
<0.20	43	2			45
≥0.20<0.50	6	35	2		43
≥0.50<0.80		11	13	1	25
≥0.80			12	23	35
Total	49	48	27	24	148

SRM indices interpreted by authors according to Cohen's thresholds for ES_{pooled}

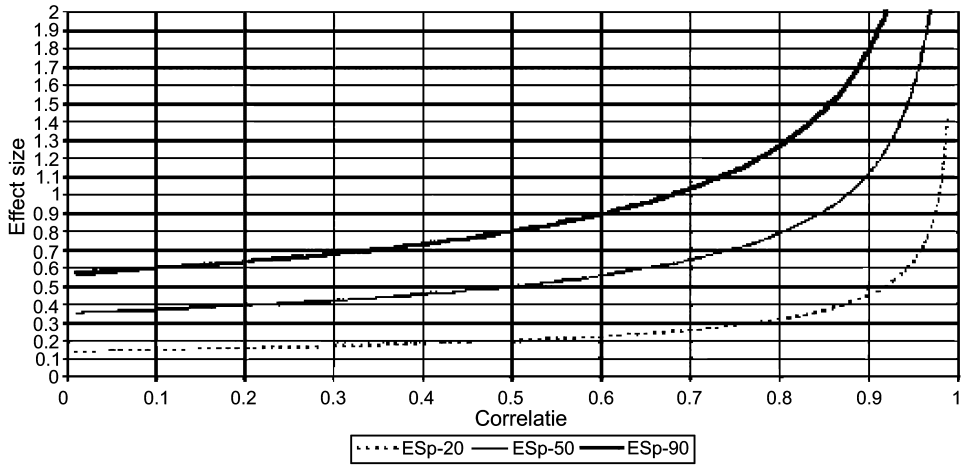

Figure 1. Cohen's threshold's for effect size SRM corrected for the size of the correllation coefficent between repeated measurements.

range of effect must be valued as 'small' with a $r = 0.14$ ($0.65/\sqrt{2}/\sqrt{1-0.14} = 0.49$).

In contrast with the fixed threshold values 0.20, 0.50 and 0.80 in Figure 1, in the analysis of 148 effect size estimates from which the correlation of a person's health status measurements over time was calculated, we found SRM values ranging from 0.04 to 2.42 [107]. Correlation coefficients ranged from 0.08 to 0.89 and 70% of the 148 coefficients were larger than 0.50. Overestimates of effect size are easily estimated. For example: an SRM of 0.85 interpreted by the researcher as large effect, changed into a moderate effect according to Cohen's thresholds, due to a correlation of 0.12 between repeated measurements

Conclusion – Discussion

Thus, the SRM interpretation of effect magnitude with the thresholds Cohen with the ES_p calculated on the same data (transformation of the same mean change over time into units of pooled standard deviation) may result in dramatic differences (23–50% of the SRM indices are overestimated). Unfortunately, we still have no algorithm for effect size indices calculated with the standard deviation from baseline scores or from change scores in stable subjects according to an external criterion. Furthermore, even in a situation where we are able to reliably interpret effect size, we cannot differentiate between a 'large' and 'very large' effect since the cut-off point for large has a theoretical range from ES 0.80 to infinite. However, Hopkins' [108] Likert-scale approach is able to give meaning to the extension of the scale to the level above large for Cohen's effect size statistic: $ES = 0- < 0.20$ trivial effect; $ES \geq 0.20- < 0.60$ small effect; $ES \geq 0.60 < 1.20$ moderate effect; $ES \geq 1.20- < 2.0$ large effect; $ES \geq 2.0- < 4.0$ very large and $ES \geq 4.0- \infty$ is considered to be 'nearly perfect'. In addition, to thresholds for effect magnitudes, Hopkins elaborated Cohen's thresholds for correlation coefficients, relative risks and odds ratio. Despite this promising attempt to proceed with a more complete scale of effect magnitude, further research will need to provide empirical evidence for the external validity of this new rule of thumb for effect size interpretation irrespective of health status measure and research designs.

Ever since Jacob Cohen wrote his well-known book [16], the effect size has been a problematic parameter in evaluation research, and several promising alternatives (for example, the ''Reliable Change Index''), have been developed [109], improved and criticised [35, 110–113]. In future studies statistical computer programmes may be able to give the researcher additional information on some intervention effect indices (notwithstanding the fact that no consensus exists on a method for signifying the magnitude of change within and between experimental and control groups that is meaningful in particular intervention contexts). Nevertheless, implementing effect sizes standard in the representation of statistical results may require researchers to change long-held patterns of behaviour.

The values used in effect size classification for difference between means as small, medium, and large was arbitrary but seemed reasonable, Cohen stated some 30 years ago. In the debate over which standardizing unit of the difference one should take in a within-group situation, we propose that estimating the magnitude of change by using either the SD of the change score or the pooled SD is preferable to the use of the SD at baseline as proposed by Kazis et al. [114], although the SRM must be adjusted to make correct use of Cohen's thresholds when magnitude of change over time is estimated in evaluation research.

These thresholds of Cohen are now being cited without distinguishing between the unit by which the assessed change over time is standardised. This is surprising since there is unequivocally no doubt that his rule of thumb was derived from the pooled SD as the estimate of the common within variance. Moreover, routine action in calculating effect sizes may have led to a reduced awareness of factors originally considered only in the calculation of power and sample size. For instance, the calculation of power of the detected change or difference without using the information of **r** can lead to the wrong inferences [16].

In evaluation research on treatment-related quality of life, researchers seem to overlook the fact that, in assessing change over time within one subject, the experimental technique of 'self-matching' reduces the proportion of the total variance due to extraneous variables not related to the treatment or intervention per se [115].

We may conclude that the rule of thumb proposed by Cohen can induce differences in the interpretation of the size of estimated effects. At present it does not appear to us that a single set of rules that is unequivocal or normative at some level is available. We have begun to explore alternative methods in effect size estimation and have assessed the interrelation between two effect sizes as estimates of magnitude of change over time within groups.

As we have demonstrated, errors can easily be made and different interpretations of the magnitude of detected change may occur. In analysing the data from our sample of published studies on change over time in health-related quality of life, we saw meaningful shifts in magnitude of detected change in relation to the size of the correlation between pre- and post-test scores. In this article we have attempted to draw the attention to the problem of over- or underestimation of effect sizes when the Standardized Response Mean is used. Studies in which the mean change over time is standardized with the $SD_{baseline}$ according to Kazis et al. [114] should report the ES_p to show that the results were not dependent on the choice of denominator in the d-index formula.

Due to their increasing appearance, it is important that all aspects of estimating the magnitude of change be inspected. One of these aspects is the consequence of the hidden role of the correlation coefficient between repeated measurements, which increases the risk of incorrect conclusions. This initial effort may provide a moderate step toward the development of a precise and useful index in quality of life assessment in clinical trials.

Recommendations for practice and research

So long as no consensus reached on standards for evaluating, using and interpreting effect size estimates of intervention-related change in evaluation research, there is an important need to develop uniform and widely accepted criteria to give meaning to the size of an effect. This lack of precision is not only relevant when evaluating intervention-related change within and between groups, but, even more important in the estimation of power in the planning phase of a trial. Standardisation of effect size interpretation needs reference ranges of health-related functional status assessed with population surveys. Furthermore, longitudinal research is needed to discriminate between changes in HRFS over time in a sample drawn from the general population, with change in a sub-sample of chronically ill patients. In other words, with knowledge about a reference range of an indicator of health-related functional status in the general population, we can recognise that there are differences. Furthermore, with a longitudinally assessed estimate of autonomous change in the same sample, we will be able to better understand the meaningfulness of intervention-related effects.

In studies on the measurement of health-related quality of life and HRFS, effect sizes (ES) have been used as surrogates for clinically relevant change when change over time in outcome was substantial. However, ES do not provide a complete understanding of the meaningfulness of the observed change. Patients have to perceive a change in the performance of daily activities in order to rate the direction and degree of change; moreover, even when this perceived change is small in magnitude, it may still be perceived as a significant one by the patient. According to Osoba [116], the significance of change as perceived by the subject 'should be of paramount consideration' in future attempts to define the meaningfulness of change in HRFS or health-related quality of life. The development of multi-item transition measures may cover change in the relevant underlying domain more representatively [107, 117]. Therefore, we suggest that measures that assess more concrete aspects of the patient's HRFS will provide greater accordance between serial and transition measures of change.

However, when a patient rates a reduction in (for example) difficulty in climbing stairs, as 'large', it does not necessarily imply that a patient will view this subjectively significant change as being important. Future areas of research aimed at quantification of meaningful change in HRFS should also include the importance patients assign to that change, even if it

is experienced as being small. One piece of research has produced examples that seem promising extensions of transition questions. In this approach, the respondent rates the direction and the degree of perceived change by a assigning a value that has meaning to the respondent for the experienced change, as well as by rating the degree of importance the respondent assigns to perceived change. In evaluation of intervention-related change in evaluation studies, the importance assigned to the small improvement in one item of a domain of HRFS may outweigh a moderate deterioration in another item belonging to the same domain.

Finally, the following are key issues in the debate on methods for estimating clinically important change: Significance of intervention effects: significance to whom [93] who is to say what is important? [90] and ''ask patients what they want'' [94, 118–120] have increasingly become apparent. To give clinically relevant meaning to change scores gained on two different points in time using HRFS instruments, several investigators suggest that the current approaches could be improved by taking more explicit account of patients' perceptions and expectations. A new paradigm is incorporating individual patient perspectives, expectations and preferences with respect to the effects of (innovative) interventions in the outcome measures. With scoring systems based on individualised measures such as the so-called Goal Attainment Scale (GAS) or Patient Specific Index (PCI), each patient essentially receives his or her 'own instrument' and these instruments seem to show an improved sensitivity to change in health-related functional status when compared with conventional methods [75, 92, 95, 121–125].

Methodological studies focussed at improving the longitudinal validity or responsiveness of health outcome measurement should be aimed at supporting, health professionals, investigators and administrators in the understanding and critical evaluation of the appropriateness of health status measures and understanding of methods in estimating and interpreting change in patient-assessed health outcomes. Health professionals increasingly stress that in the realisation of effective care and expected outcome of planned change in the process of care delivery, patients' preferences are essential sources of information. The operationalisation of the patient's perception of the severity of limitation in domains of health-related functioning, or operationalisation of individual preference or weighted relevance of items of health-related functional status measures is still in its infancy. However, for health administrators and decision-makers, investigation into the validity of patient-specific HRFS instruments used to evaluate the outcomes of innovative and care, standardisation of methods is required. HRFS instruments cannot be used in the evaluation of treatment and care without a valid way of ascertaining what change in measured difference scores means.

Acknowledgments

Appreciation is expressed to Drs. Roy Stewart for providing valuable assistance with several aspects of the analysis, and a critical review of the manuscript. Prof. Dr. Wim van den Heuvel and Dr. Mike de Jongste provided helpful reviews of the manuscript.

References

1. Testa MA, Nackley JF. Methods for quality-of-life studies. Annual Review of Public Health 1994;15:535–59.
2. Hunt SM. The problem of quality of life. Quality of Life Research 1997;6:205–12.
3. Gill TM, Feinstein AR. A critical appraisal of the quality of quality of life measurements. Journal of the American Medical Association 1994;272:619–26.
4. Spilker B. Quality of life and Pharmacoeconomics in clinical trials. 2nd ed. Philadelphia: Lippincott-Raven; 1996.
5. Browne JP, McGee HM, O'Boyle CA. Conceptual approaches to the assessment of quality of life. Psychology and Health 1997;12(6):737–51.
6. Bonomi AE, Patrick DL, Bushnell DM, Martin M. Quality of life measurement. Will we ever be satisfied? Journal of Clinical Epidemiology 2000;53(1):19–23.
7. Anderson KL, Burckhardt CS. Conceptualization and measurement of quality of life as an outcome variable for health care intervention and research. Journal of Advanced Nursing 1999;29(2):298–306.
8. Fitzpatrick R. A pragmatic defence of health status measures. Health Care Analysis 1996;4:265–72.
9. Kempen GIJM, Steverink N, Ormel J, Deeg DJH. The assessment of ADL among frail elderly in an interview survey: self-report versus performance-based tests and determinants of discrepancies. Journal of Gerontology: Psychological Sciences 1996;51B(5):254–60.
10. Van Heuvelen MJG. Physical activity, physical fitness and disability in older persons. (Dissertation). Groningen: Rijksuniversiteit Groningen; 1999.
11. Emery CF, Blumenthal JA. Perceived change among participants in an exercise program for older adults. The Gerontologist 1990;30(4):516–21.

12. Rosendal H, van Beekum WT, Nijhof P, De Witte LP, Schrijvers AJP. Can shared care deliver better outcomes for patients undergoing total hip replacement? A prospective assessment of patient outcomes and associated service use. International Journal of Integrated care [serial online] 2000 Nov 1;1. Available from: URL:http://www.ijic.org/.
13. Terwee CB. Graves' opthalmopathy through the eyes of the patient. Assessment of health-related quality of life. University of Amsterdam; 2000.
14. Rosendal H, Wolters CAM, Beusmans GHMI, Witte LP, Boiten J, Crebolder HFJM. Stroke service in the Netherlands: an exploratory study on effectiveness, patient satisfaction and utilisation of healthcare. International Journal of Integrated care [serial online] 2002 Mar 1;2. Available from: URL:http://www/ijic.org/.
15. Cohen J. The earth is round ($p < 0.05$). American Psychologist 1994;49(12):997–1003.
16. Cohen J. Statistical power analysis for the behavioural sciences. rev. ed. New York: Academic Press; 1977.
17. Stockler MR, Osoba D, Goodwin P, Corey P, Tannock IF. Responsiveness to change in health-related quality of life in a randomized clinical trial: a comparison of the prostate cancer specific quality of life instrument (PROSQOLI) with analogous scales from the EORTC QLQ-C30 and a trial specific module. Journal of Clinical Epidemiology 1998;51(2):137–45.
18. Murawski MM, Miederhoff PA. On the generalizability of statistical expressions of health related quality of life instrument responsiveness: a data synthesis. Quality of Life Research 1998;7:11–22.
19. Taylor R, Kirby B, Burdon D, Caves R. The assessment of recovery in patients after myocardial infarction using three generic quality-of-life measures. Journal of Cardiopulmonary Rehabilitation 1998;18:139–44.
20. Wiebe S, Rose K, Derry P, McLachlan R. Outcome assessment in epilepsy: comparative responsiveness of quality of life and psychosocial instruments. Epilepsia 1997;38(4):430–8.
21. Russel MGVM, Pastoor CJ, Brandon S, Rijken J, Engels LGJB, Van der Heijde DMFM, et al. Validation of the Dutch translation of the inflammatory bowel disease questionnaire (IBDQ): a health related quality of life questionnaire in inflammatory bowel disease. Digestion 1997;58:282–8.
22. Tuley MR, Mulrow CD, McMahan CA. Estimating and testing an index of responsiveness and the relationship of the index to power. Journal of Clinical Epidemiology 1991;44(4/5):417–21.
23. Parkerson GR, Willke RJ, Hays RD. An International comparison of the reliability and responsiveness of the Duke health profile for measuring health-related quality of life of patients treated with Alprostadil for erectile dysfunction. Medical Care 1999;37(1):56–67.
24. Wasserfallen JB, Gold K, Schulman KA, Baraniuk JN. Development and validation of a rhinoconjunctivitis and asthma symptom score for use as an outcome measure in clinical trials. Journal of Allergy and Clinical Immunology 1997;100(1):16–22.
25. Guyatt GH, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. Journal of Chronical Disease 1987;40(2):171–8.
26. Jenkinson C, Lawrence K, McWhinnie D, Gordon J. Sensitivity to change of health status measures in a randomized controlled trial: comparison of COOP charts and the SF-36. Quality of Life Research 1995;4:47–52.
27. Katz JN, Larson MG, Phillips CB, Fossel AH, Liang MH. Comparative measurement sensitivity of short and longer health status instruments. Medical Care 1992;30(10):917–25.
28. Keoghane SR, Lawrence KC, Jenkinson CP, Doll HA, Chappel DB, Cranston DW. The Oxford Laser Prostate Trial: sensitivity to change of three measures of outcome. Urology 1996;47(1):43–47.
29. Garratt AM, Ruta DA, Abdalla MI, Russell T. Responsiveness of the SF-36 and a condition-specific measure of health for patients with varicose veins. Quality of Life Research 1996;(5):223–34.
30. Bousquet J, Duchateau J, Pignat JC, Fayol C, Marquis P, Mariz S, et al. Improvement of quality of life by treatment with cetirizine in patients with perennial allergic rhinitis as determined by a French version of the SF-36 questionnaire. Journal of Allergy and Clinical Immunology 1997;98(2):309–16.
31. Jenkinson C, Layte R, Coulter A, Wright L. Evidence for the sensitivity of the SF-36 health status measure to inequalities in health: results from the Oxford healthy lifestyles survey. Journal of Epidemiology and Community Health 1996;(50):377–80.
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. Journal of Clinical Epidemiology 1996;49(7):711–7.
33. Andresen EM, Bowley N, Rothenberg BM, Panzer R, Katz P. Test-retest performance of a mailed version of the medical outcomes study 36-item short-form health survey among older adults. Medical Care 1996;34(12):1165–70.
34. Ziebland S. The short form 36 health status questionnaire: clues from the Oxford region's normative data about its usefulness in measuring health gain in population surveys. Journal of Epidemiology and Community Health 1995;(49):102–5.
35. Speer DC. Clinically significant change: Jacobson and Truax (1991) revisited [published erratum appears in Journal of Consulting and Clinical Psychology 1993 Feb;61(1):27). Journal of Consulting and Clinical Psychology 1992;60(3):402–8.
36. Juniper EF. Measuring health-related quality of life in rhinitis. Journal of Allergy and Clinical Immunology 1997;99(2):S742–9.

37. de Bruin AF, Diederiks JPM, De Witte LP, Stevens FCJ, Philipsen H. Assessing the responsiveness of a functional status measure: the sickness impact profile versus the SIP68. Journal of Clinical Epidemiology 1997;50(5):529–40.
38. de Bruin AF. The measurement of sickness impact; the construction of the SIP68 (dissertation). Maastricht: Rijksuniversiteit Limburg; 1996.
39. Jenkinson C, Layte R, Jenkinson D, Lawrence K, Petersen S, Paice C, et al. A shorter form health survey: can the SF-12 replicate results from the SF-36 in longitudinal studies? Journal of Public Health Medicine 1997;19(2):179–86.
40. Roberts R, Hemingway H, Marmot M. Psychometric and clinical validity of the SF-36 General Health Survey in the Whitehall II study. British Journal of Health Psychology 1997;(2):285–300.
41. International Resource Center for Health Care Assessment (IRC). How to score the MOS 36-item Short Form Health Survey (SF36): SF-36TM scoring rules. Boston: New England Medical Center Hospitals; 1991.
42. Stadnyk K, Calder J, Rockwood K. Testing the measurement properties of the Short Form-36 Health Survey in a frail elderly population. Journal of Clinical Epidemiology 1998;51(10):827–35.
43. Jenkinson C, Layte R, Lawrence K. Development and testing of the medical outcomes study 36-item Short Form Health Survey summary scale scores in the United Kingdom. Results from a large-scale survey and a clinical trial. Medical Care 1997;35(4):410–6.
44. McHorney CA, Ware JEJ, Raczek AE. The MOS 36-Item short-form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. Medical Care 1993;31(3):247–63.
45. Jenkinson C, Coulter A, Wright L. Short form 36 (SF36) health survey questionnaire: normative data for adults of working age [see comments]. British Medical Journal 1993;306(6890):1437–40.
46. Van der Zee K, Sanderman R, Heyink JW, De Haes H. Psychometric qualities of the RAND 36-item Health Survey 1.0: a multidimensional measure of general health status. International Journal of Behavioral Medicine 1996;(3):104–22.
47. Hawker G, Melfi C, Paul J, Green R, Bombardier C. Comparison of a generic (SF-36) and a disease-specific (WOMAC) instrument in the measurement of outcomes after knee replacement surgery. Journal of Rheumatology 1995;22:1193–6.
48. Ware JE, Sherbourne CD. The MOS 36-item Short-Form Health Survey (SF-36): I. Conceptual framework and item selection. Medical Care 1992;30(473):483.
49. Van der Zee K, Sanderman R. Het meten van de algemene gezondheidstoestand met de RAND-36, een handleiding [The assessment of general health status with the RAND-36]. Noordelijk Centrum voor Gezondheidsvraagstukken; 1993.
50. Van der Zee K, Sanderman R, Heyink JW. De psychometrische kwaliteiten van de MOS Short Form health Survey (SF-36) in een Nederlandse populatie [The psychometric properties of the SF-36 in a Dutch population]. Tijdschrift voor Sociale Gezondheidszorg 1993;71:183–91.
51. Wells G, Boers M, Shea B, Tugwell P, Westhovens R, Saurez-Almazor M, et al. Sensitivity to change of generic quality of life instruments in patients with rheumatoid arthritis: preliminary findings in the generic health OMERACT Study. The Journal of Rheumatology 1999;26(1):217–21.
52. Ware JE, Kosinski M, Bayliss MS, McHorney CA, Rogers WH, Raczek AE. Comparison of methods for the scoring and statistical analysis of SF-36 health profiles and summary measures: summary of results from the Medical Outcome Study. Medical Care 1995;33(Suppl. 4):AS264–79.
53. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. Controlled Clinical Trials 1991;12(4 Suppl):S142–58.
54. Katz JN, Gelberman RH, Wright EA, Lew RA, Liang MH. Responsiveness of self-reported and objective measures of disease severity in Carpal Tunnel Syndrome. Medical Care 1994;32(11):1127–33.
55. Middel B, Kuipers-Upmeijer H, Bouma J, Staal MJ, Oenema D, Postma Th, et al. Effect of intrathecal baclofen delivered by an implanted programmable pump on health related quality of life in patients with severe spasticity. Journal of Neurology Neurosurgery and Psychiatry 1997;63:204–9.
56. de Beurs E, van Balkom AJLM, Lange A, Koele P, van Dyck R. Treatment of panic disorder with agoraphobia: comparison of fluvoxamine, placebo, and psychological panic management combined with exposure and of exposure in vivo alone. American Journal of Psychiatry 1995;152(5):683–91.
57. Sneeuw KCA, Aaronson NK, Sprangers MAG, Detmar SB, Wever LDV, Schornagel JH. Comparison of patient and proxy EORTC QLQ-C30 ratings in assessing the quality of life of cancer patients. Journal of Clinical Epidemiology 1998;51(7):617–31.
58. Van der Windt DAWM, Van der Heijden GJMG, De Winter AF, Koes BW, Deville W, Bouter LM. The responsiveness of the shoulder disability questionnaire. Annals of the Rheumatic Diseases 1998;57:82–87.
59. Bain BA, Dollaghan CA. Clinical Forum: treatment efficacy. The notion of clinically significant change. Language, Speech, and Hearing in Schools 1991;22:264–70.
60. Osoba D, Rodrigues G, Myles J, Zee B, Pater J. Interpreting the significance of changes in health-related quality-of-life scores. Journal of Clinical Oncology 1998;16(1):139–44.
61. Husted JA, Cook RJ, Farewell VTGDD. Methods for assessing responsiveness: a critical review and recommendations. Journal of Clinical Epidemiology 2000;53:459–68.
62. Ziebland S. Measuring changes in health status. In: Jenkinson C, editor. Measuring health and medical outcomes. London: UCL Press; 1994. p. 42–53.

63. Ziebland S, Fitzpatrick R, Jenkinson C, Mowat A, Mowat A. Comparison of two approaches to measuring change in health status in rheumatoid arthritis: the health assessment questionnaire (HAQ) and modified HAQ. Annals of the Rheumatic Diseases 1992;(51):1202–5.
64. Stucki G, Liang MH, Fossel AH, Katz JN. Relative responsiveness of condition-specific and generic health status measures in degenerative lumbar spinal stenosis. Journal of Clinical Epidemiology 1995;48(11):1369–78.
65. Vliet-Vlieland ThPM, Zwinderman AH, Breedveld FC, Hazes JMW. Measurement of morning stiffness in rheumatoid arthritis clinical trials. Journal of Clinical Epidemiology 1997;50(7):757–63.
66. Guyatt GH, Kirshner B, Jaeschke R. Measuring health status: what are the necessary measurement properties? Journal of Clinical Epidemiology 1992;45(12):1341–5.
67. Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. Annals of Internal Medicine 1993;118(8):622–9.
68. Norman G. Issues in the use of change scores in randomized trials. Journal of Clinical Epidemiology 1989;42(11):1097–1105.
69. Stewart AL, Greenfield S, Hays RD, Wells K, Rogers WH, Berry SD, et al. Functional status and well-being of patients with chronic conditions: results from the medical outcome study. Journal of the American Medical Association 1989;262(7):907–13.
70. Patrick DL, Deyo RA. Generic and disease-specific measures in assessing health status and quality of life. Medical Care 1989;27S:S217–32.
71. Middel B, Bouma J, Crijns HJGM, De Jongste MJL, Van Sonderen FLP, Niemeijer MG, et al. The psychometric properties of the Minnesota Living with Heart Failure Questionnaire (MLHF-Q). Clinical Rehabilitation 2001;15:380–91.
72. Hawley DJ, Wolfe F. Sensitivity to change of the health assessment questionnaire (HAQ) and other clinical and health status measures in rheumatoid arthritis: results of short-term clinical trials and observational studies versus long-term observational studies [published erratum appears in Arthritis Care and Research 1992 Dec;5(4):229]. Arthritis Care and Research 1992;5(3):130–6.
73. Gliklich RE, Hilinsky JM. Longitudinal sensitivity of generic and specific health measures in chronic sinusitis. Quality of Life Research 1995;4:27–32.
74. Wright JG, Young NL. A comparison of different indices of responsiveness. Journal of Clinical Epidemiology 1997:50(3):239-46.
75. Bessette L, Sangha O, Kuntz KM, Keller RB, Lew RA, Fossel AH, et al. Comparative responsiveness of generic versus disease-specific and weighted versus unweighted health status measures in carpal tunnel syndrome. Medical Care 1998;36(4):491–502.
76. Vaile JH, Mathers M, Ramos-Remus C, Russel AS. Generic health instruments do not comprehensively capture patient perceived improvements in patients with carpal tunnel syndrome. The Journal of Rheumatology 1999;26(5):1163–6.
77. Rosnow RL, Rosenthal R. Statistical procedures and the justification of knowledge in psychological science. American Psychologist 1989;44:1276–84.
78. Rosenthal R. Progress in clinical psychology: Is there any? Clinical Psychology: Science and Practice 1995;2:133–50.
79. Rosenthal R, Rubin DB. The Counternull value of an effect size: a new statistic. Psychological Science 1994;5(6):329–34.
80. Bartko JJ, Pulver AE, Carpenter WT. The power of analysis: statistical perspectives. Part 2. Psychiatry Research 1988;23:301–9.
81. Rozeboom WW. The fallacy of the null-hypothesis significance test. Psychological Bulletin 1960;57:416–28.
82. Cohen J. Things I have learned (so far). American Psychogist 1992;45:1304–12.
83. Levin JR, Robinson DH. Further reflections on hypothesis testing and editorial policy for primary research journals. Educational Psychology Review 1999;11(2):143–55.
84. Thompson B. If statistical significance tests are broken/misused, what practices should supplement or replace them? Theory & Psychology 1999;9(2):165–81.
85. Tukey JW. The philosophy of multiple comparisons. Statistical Science 1991;6:100–16.
86. Thompson B. Editorial policies regarding statistical significance tests: futher comments. Educational Research 1997;26(5):29–32.
87. Murphy KR. Editorial. Journal of Applied Psychology 1997;82:3–5.
88. Kirk RE. Practical significance: a concept whose time has come. Educational and Psychological Measurement 1996;56(5):746–59.
89. Naylor CD, Llewellyn-Thomas HA. Can there be a more patient-centered approach to determining clinically important effect sizes for randomized treatment trials? Journal of Clinical Epidemiology 1994;47(7):787–95.
90. Lachs MS. The more things change... Journal of Clinical Epidemiology 1993;46(10):1091–2.
91. Wright JG. The minimal important difference: Who's to say what is important? Journal of Clinical Epidemiology 1996;49(11):1221–2.
92. Tugwell P, Bombardier C, Buchanan WW, Goldsmith CH, Grace E, Hanna B. The MACTAR patient preference disability questionnaire – an individualized functional priority approach for assessing improvement in physical disability in clinical trials in rheumatoid arthritis. Journal of Rheumatology 1987;14(3):446–51.

93. Mitchell PH. The significance of treatment effects: significance to whom? Medical Care 1995;33(4):AS280–5.
94. Wright JG, Rudicel S, Feinstein AR. Ask patients what they want. Evaluation of individual complaints before total hip replacement. Journal of Bone and Joint Surgery 1994;76-B(2):229–34.
95. Rockwood K, Stolee P, Fox RA. Use of goal attainment scaling in measuring clinically important change in the frail elderly [see comments]. Journal of Clinical Epidemiology 1993;46(10):1113–8.
96. Leon AC, Shear K, Portera L, Klerman GL. Effect size as a measure of symptom-specific drug change in clinical trials. Psychopharmacology Bulletin 1993;29(2):163–7.
97. Pulver AE, Bartko JJ, McGrath JA. The power of analysis: statistical perspectives. Part 1. Psychiatry Research 1988;23:295–9.
98. Brewer JK. Effect size: the most troublesome of the hypothesis testing considerations. CEDR Quarterly 1978;11(4):7–10.
99. Borenstein M. A note on the use of confidence intervals in psychiatric research. Psychopharmacology Bulletin 1994;30(2):235–8.
100. Cooper HM. On the significance of effects and the effects of significance. Journal of Personality and Social Psychology 1981;41(5):1013–8.
101. van Bennekom CAM, Jelles F, Lankhorst GJ, Bouter LM. Responsiveness of the rehabilitation activities profile and the Barthel Index. Journal of Cliical Epidemiology 1996;49(1):39–44.
102. Beurskens AJHM, de Vet HCW, Koke AJA. Responsiveness of functional status in low back pain: a comparison of different instruments. Pain 1996;65:71–6.
103. Kempen GIJM, Miedema I, van den Bos GAM, Ormel J. Relationship of domain-specific measures of health to perceived overall health among older subjects. Journal of Clinical Epidemiology 1998;51(1):11–18.
104. Kraemer HC. Reporting the size of effects in research studies to facilitate assessment of practical or clinical significance. Psychoneuroendocrinology 1992;17(6):527–36.
105. Lipsey MW. Design sensitivity. Statistical power for experimental research. London: SAGE Publications; 1990.
106. Liang MH, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopedic evaluation. Medical Care 1990;28(7):632–42.
107. Middel B. Assessment of change in clinical evaluation. Northern Centre for Healthcare Research. The Netherlands: University of Groningen; 2001.
108. Hopkins WG. A new view of statistics: Effect Magnitudes. 1997. Available from: URL:http://sportsci.org/resource/stats/effectmag.html.
109. Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. Journal of Consulting and Clinical Psychology 1991;59:12–19.
110. Hageman WJJM, Arrindell WA. A further refinement of the reliable change (RC) index by improving the pre-post difference score: introducing RCID. Behaviour Research and Therapy 1993;31(7):693–700.
111. Hageman WJ, Arrindell WA. Establishing clinically significant change: increment of precision and the distinction between individual and group level of analysis [see comments]. Behaviour Research and Therapy 1999;37(12):1169–93.
112. Hageman WJ, Arrindell WA. Clinically significant and practical! Enhancing precision does make a difference. Reply to McGlinchey and Jacobson, Hsu, and Speer. Behaviour Research and Therapy 1999;37:1219–33.
113. Maassen GH. Kelley's formula as a basis for the assessment of reliable change. Psychometrika 2000;65(2):187–97.
114. Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. Medical Care 1989;27(3, Suppl):S178–189.
115. Winer BJ. Statistical principles in experimental design. 2nd ed. Tokyo: McGraw-Hill Kogakusha; 1962.
116. Osoba D. Interpreting the meaningfulness of change in health-related quality of life scores: lessons from studies in adults. International Journal of Cancer 1999;12:132–7.
117. Middel B, de Greef M, Crijns HJGM, De Jongste MJL, Stewart R, van den Heuvel WJA. Why don't we ask patients with heart failure directly how much they have changed after treatment? A comparison of retrospective multi-item scales with serial change in domains of health-related functional status. Journal of Cardiopulmonary Rehabilitation 2002;22(11):47–52.
118. Liang MH, Larson MG, Cullen KE, Schwartz JA. Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. Arthritis and Rheumatism 1985;28(5):542–7.
119. Cassard SD, Patrick DL, Damiano AM, Legro MW, Tielsch JM, Diener West M, et al. Reproducibility and responsiveness of the VF-14. An index of functional impairment in patients with cataracts. Archives of Ophthalmology 1995;113(12):1508–13.
120. Wright JG, Young NL. The patient-specific index: asking patients what they want. The Journal of Bone and Joint Surgery 1997;79-A(7):974–83.
121. Tugwell P, Bombardier C, Buchanan WW, Goldsmith C, Grace E, Bennett KJ, et al. Methotrexate in rheumatoid arthritis. Impact on quality of life assessed by traditional standard-item and individualized patient preference health status questionnaires. Archives of Internal Medicine 1990;150:59–62.
122. MacKenzie RC, Charlson ME, DiGioia D, Kelley K. A patient-specific measure of change in maximal function. Archives of Internal Medicine 1986;146:1325–9.

123. Rockwood K, Joyce B, Stolee P. Use of goal attainment scaling in measuring clinically important change in cognitive rehabilitation patients. Journal of Clinical Epidemiology 1997;50(5):581–8.
124. Gordon JE, Powell C, Rockwood K. Goal attainment scaling as a measure of clinically important change in nursing-home patients. Age and Ageing 1999;28:275–81.
125. Gordon J, Rockwood K, Powell C. Assessing patients' views of clinical changes [letter]. Journal of the American Medical Association 2000;283(14):1824–5.