6

Why don't we ask patients with coronary artery disease directly how much they have changed after treatment? A comparison of retrospective multi-item change scales with serial change in domains of health related functional status.

Berrie Middel, Msc*, Eric van Sonderen, PhD*, Mathieu de Greef, PhD*, Harry, J.G.M. Crijns, MD,PHD+, Mike J.L. de Jongste, MD, PhD+, Roy Stewart, Msc*, Wim J.A. van den Heuvel, PhD*.

* Northern Centre for Healtcare Research. School of Medicine, University of Groningen, The Netherlands

\+ Department of Cardiology, Thoraxcenter, University Hospital of groningen, The Netherlands

Submitted

ABSTRACT

Purpose
In the literature on the measurement of treatment-related change some authors advocate the use of change scores as the best approach; others prefer the approach to measure change by asking patients how much they have changed after treatment. The aim of this study was to compare these two methods in measurement of changes in identical domains of health –related functional status (HRFS).

Methods.
Analogous to the widely used single global or transition questions which assess the patient's retrospectively perceived change, we modified every item belonging to domains of health of the Minnesota Living with Heart Failure-Questionnaire (MLHF-Q), into transition questions. Subjects were 217 patients with heart failure, who underwent a treatment with known efficacy, were selected.

Results.
The reliability of the scale 'physical functioning' as the composite of change scores, and of the concordant scale composed of identical transition items were estimated and yielded a Cronbach's alpha of 0.86 and 0.92, respectively, whereas the change in 'emotional functioning' and the concordant transition scale yielded identical alphas of 0.76. Factor analysis yielded similar and clearly interpretable dimensions for both methods. The canonical correlation (R_c) between the composite of the change score items and the composite of the concordant transition items that belong to the domain of physical functioning was R_c=.63 ($p < .001$), whereas the combination of the emotional functioning-items yielded a R_c=.48 ($p< .001$) with a percentage of linearly explained variance between the two dimensions (40% and 23%, respectively).

Conclusions.
The patient's retrospective assessment of change after intervention appears to provide reliable and valid information compared to prospective change scores when items are used beloning to validated measures health-status.

6.1 INTRODUCTION

Clinicians and other health professionals have difficulty to find Health-Related Functional Status (HRFS) measures clinical important change that is expected to occur after treatment evaluated in cardiological trials. To evaluate clinically relevant change reliable and valid, the patient's perception or opinion of what constitutes relevant change in health status domains must be captured in a multi-dimensional mode. However, treatment-related change over time is usually based on change or difference scores that are calculated simply by subtracting baseline scores from post-treatment scores. The interpretation of the amount of change over time as clinically relevant depends on a subjective judgement of either clinician or patients as a substitute for a golden standard. Therefore, a widely accepted solution to discriminate between relevant and irrelevant change is the use of so-called 'transition' or 'global questions' as external criterion or standard, by asking the patients retrospectively how much they feel better or worse compared to the situation at baseline.

These transition questions are used in several modes: 1) single, retrospective questions at follow-up about the direction and magnitude of perceived change in general or 2) after treatment in domains of health, e.g. physical, emotional and social functioning, 3) about satisfaction with the change after treatment, 4) perceived degree of change in difficulty to accomplish a specified task or 5) as a serially assessed retrospective global rating to examine incremental perceived change between baseline and follow-up [1-12]. Another reason to use transition questions is to compare the results of assessing change in health-related quality of life with both repeated measurement and with the patient's retrospective perception of change after treatment.

In several studies, the accuracy, precision, reliability and validity of *single* global or transition questions of health have been discussed [6,13-19]. The main disadvantage of a single item of retrospectively perceived change in overall health is that the answer "since the operation my health state has worsened" does not cover domain-specific change in health. We can imagine that the improvement in the domain of physical health is overshadowed by the perception of a worsening in emotional functioning and in such situations a global single transition question is not a valid indicator of change in health status. Another disadvantage of a single question used to capture perceived change in specific domains of the patient's life (physical, emotional or social functioning) is that the internal consistency or reliability cannot be estimated. Therefore, we have good reasons to presume that multiple-item transitional scales tend to be more reliable than single-items [20] and, when these retrospective measures are conceptually identical with the longitudinally assessed change in health domains,

they would also have a better validity [21]. On the one hand, several studies have provided evidence that change in health-status domains estimated with direct transition questions were more accurate when compared with estimates derived from change scores [6,11,16,17]. On the other hand, the direction of prospective or serial change does not always correspond with the content of the single transition question(s) and therefore may miss some contrary changes [11,21,22]. To solve this, it has been suggested that patients should be asked a series of transition judgements about functional limitations related to their disease or to areas of health status such as physical, emotional and social function [10,17]. However, we have not been able to find studies in which retrospective change in domains of health was measured with multi-transition items intended to contribute to the assessment of change in an underlying dimension of health status. Two studies we have found used a set of multiple transition items but without a relationship with an underlying dimension of health [11,23]. No study was found in which an attempt was made to investigate the question whether both methods measure the same dimensions of health status by means of the summed composite of item scores (scales). Also no study was found in which the concordance between the perceived change and change scores, calculated in a before-after design, was evaluated simultaneously in a one-to one relationship with items and with scales. Before we can reasonably claim that multiple item transition scales measure change in health-related functional status (HRFS), we have to demonstrate that this new operational procedure yields substantially the same results as measuring change with the standard repeated measurement procedure. To this end, his paper examines the reliability, the convergent validity and 'known groups' validity of the patient's view of change in domains of health as outcome measures. The following questions are addressed:

1. Do we measure change in the same domains of HRFS if we use multi-item transition scales assessing (retrospective) perceived change compared with (prospective) change assessed with repeated measures?
2. To what extent does concordance exist between health status scales composed of prospective change scores and those composed of retrospective transition scores?
3. How do both instruments compare in their ability to discriminate between groups known to differ on a measure external to the questionnaire (i.e. disappearance of angina pectoris)?

6.2 METHODS

To ensure that change in health status occurred we selected a group of patients undergoing a treatment with known efficacy and selected a disease-specific instrument with known sensitivity to detect change over time [24-26]. The study sample was composed of two groups: 1. patients who, after a diagnostic Coronary Angio Graphy (CAG), were scheduled for an invasive treatment Percutaneous Transluminal Coronary Angioplasty (PTCA) or Coronary Artery Bypass Grafting (CABG) and 2. patients who needed no operative intervention after CAG and were treated with pharmacotherapy.
The aim of the current study was to compare potentially similar strategies of measuring treatment-related change in two well-defined important domains of HRFS: physical functioning, and emotional functioning. Consequently, items that are not purported to measure either the domain of physical or emotional functioning are not used in the comparison. To assess serial change in the domain of physical functioning we used the 8-item scale from a disease-sepcific HRFS instrument (namely the Minnesota Living with Heart Failure Questionnaire (MLHF-Q) [27,28] and 3 items from the MOS-20. [29] To assess serial change in the domain of emotional functioning the 5-item scale of the MLHF-Q was used. To assess perceived change in these domains of health with a direct method we modified the selected items of the MLHF-Q physical functioning scale and MLHF-Q emotional functioning scale and the MOS-20 items into direct questions (transition questions). These multi-item transition scales were administered at post-treatment.

6.2.1. Patient selection

Consecutive patients who, following a Coronary Angiography (CAG), were scheduled for Percuteneous Transluminal Coronary Angioplasty (PTCA) or Coronary Artery Bypass Grafting(CABG) or who needed no operative intervention but were treated with a (modified) pharmaco-therapy, were recruited from January to December 1998 from the Groningen University Hospital, the Martini Hospital, Groningen, and the Weezenlanden Hospital in Zwolle in the Netherlands. Patients with other incapacitating diseases, with cognitive impairments, aged 75 or older, or who did not speak Dutch were excluded. Ethical approval was obtained from the ethics committee at each participating hospital.

After inclusion patients received a mailed questionnaire accompanied by a written informed consent form. The questionnaire was serially administered at baseline and 6 weeks after the decision for non-invasive intervention or 6 weeks after the day a PTCA /CABG-intervention was executed. After the questionnaires were received, they were routinely checked on completeness at baseline as well as at follow-up. If questions or pages had not been filled in, either a copy was sent with a kind request to complete the questions or, in the cases of it being one question, patients were interviewed by telephone. Because the completeness of the questionnaire was monitored by a computer-programme both at baseline and follow up, we effectively reduced the non-response on questions, and consequently, on scales.
To ascertain the assessment of substantial treatment-related change we approached patients treated with interventions with known efficacy, i.e. invasive treatments PTCA or CABG and non-invasive pharmacotherapy. We presumed that at baseline i.e. prior to CAG, both patients and cardiologists had no information about decision concerning either intervention and would not affect the assessment of subjective health and should reduce the risk of floor and ceiling effects. However, this control for potential bias resulted in logistic problems and, six months after the start of the study, we were forced to select patients waiting for outpatient treatment (PTCA) or waiting for hospital admission (CABG) somewhat later after the decision was taken.

6.2.2. Data Collection and measures

The Minnesota Living with Heart Failure Questionnaire (MLHF-Q) is a disease-specific instrument which is composed of 21 items and three scales that measure the following: the physical functioning dimension (8 items), the emotional functioning dimension (5 items) and the overall score on health-related quality of life (21 items). Eight separate items do not assess an underlying dimension of health-related quality of life and therefore were not used for the current paper. These eight items measure several meaningful social and economic impairments that patients relate to their heart failure, although these 'socio-economic' items are used as a part of the overall score [27,30-33]. However, one item from the MLHF-Q had no correlation with the physical functioning scale, as predefined by Rector et al [28] both in a previous Dutch sample [26] and in the current study. Therefore, the item "'did your heart failure prevent you from living as you wanted by making your relating to or doing things with your friends or family difficult?" was skipped for scale construction and not used in further analysis. Finally, both the items from the MLHF-Q and the MOS-20 (10 items) were used in the analysis of the concordance between two methods of measuring change in the domain of physical functioning.

The response options range from "no" (score 0); very little (score 1) to very much (score 5). The total score of, the physical dimension (sub-scale) ranges from 0 to 40, the emotional dimension (sub-scale) ranges from 0 to 25. To investigate whether differences between the MLHF-Q and a generic measure of physical functioning would exist we have extended the questionnaire with 3 items from the MOS-20 [29] but with the response options analogue to the questionnaire's format. These three items had the following format: "Did your heart failure prevent you from living as you wanted **during the last month** by making it difficult for you 1) to bend, stoop or lift light objects?, 2) lift heavy objects, like moving a table? And 3) run at a fast pace? "

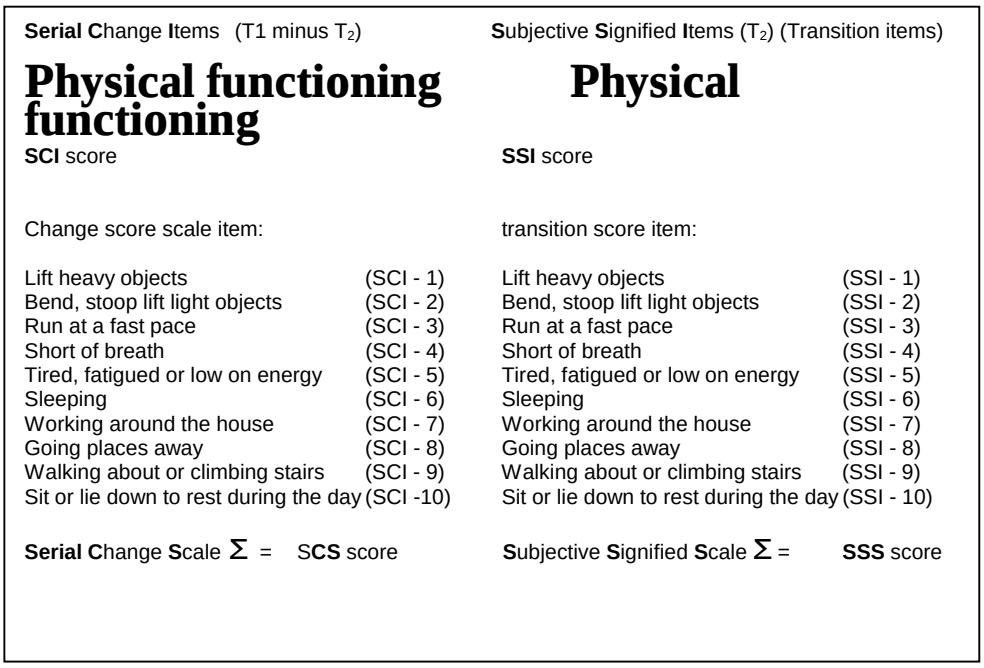
Two methods of assessing change in health-related quality of life (HRQL) with multi-item scales were applied with the study data: The first method, repeated baseline measurement of HRQL-domains with items from the MLHF-Q and MOS-20. In addition, the repeatedly measured battery of questions was transformed in a retrospective 'transition question' mode. Consequently, the patient's perception or subjective significance of change was captured at follow-up of each of the MLHF-Q and MOS-20 items in terms of the extent of feeling improved, deteriorated or not changed. Hence, the degree to which they perceived that change had occurred, on that particular item was rated on a 7-point Likert scale. Following Osoba et al. [15], we have chosen to use the term "subjective significance" because it indicates whose judgement was used to determine the direction and magnitude of change. These transition items are designed to elicit information regarding perceived change over time in specific aspects belonging to domains of HRFS.

Figure 6.1 shows, with the physical functioning items, these two strategies of assessing change in HRFS:

1. The original MLHF-Q items were phrased as follows: "Did your heart failure prevent you from living as you wanted **during the last month** by making your sleeping well at night difficult?" Serial change scores on items (SCI-scores) were calculated by subtracting the follow-up score from the baseline score to get positive numerical change data indicating improvement and negative numerical change data indicating deterioration. A change score of zero was considered to indicate neither improvement nor deterioration. With the summed composite of the SCI scores we constructed the serial change scale (SCS).
2. The Subjective Signified Items (SSI) questionnaire was used to classify patients according to whether they had improved or deteriorated on each item of the questionnaire belonging to the dimension of physical and emotional functioning. The questions were phrased as follows: "Since the last time I filled out the questionnaire (or: since my operation), my problems with walking about or climbing stairs related to my heart failure have become.

3. For every SSI item each patient was asked to circle the answer that best described the perception of the direction and magnitude of change at follow-up on a 7-point Likert scale: 1) a great deal worse; 2) moderately worse; 3) a little worse; 4) no change; 5) a little better; 6) moderately better and 7) a great deal better. With the summed composite of the SSI scores we constructed the Subjective Signified Scale (SSS). Figure 6.1 shows a general representation of prospective and retrospective methods with the physical functioning scale items.

Figure 6.1 *General representation of the prospective and retrospective method of assessing change in physical functioning.*



Functional impairment due to angina was assessed using a set of questions corresponding to the Canadian Cardiovascular Society (CCS), and to the New York Heart Association (NYHA) respectively.

6.2.3. Statistical analysis

To investigate whether the prescribed underlying domains of health of the baseline scale-items by Rector et al. [28] were measured with the same results with the change score items and the transition items, LISREL analysis was used to test the equality of factor structures (principal component analysis). Cronbach's α was used to examine the internal consistency of the transition scales that emerged from these analyses. Canonical correlation analysis was applied as a general procedure for investigating the relationships between two sets of variables. With canonical correlation analysis we transformed the prospective change item scores (PCI scores) from the set of, for example, physical functioning items into a linear combination, which is called the canonical variable. The linear combination, or canonical variable, was also constructed from the concordant set of subjective signified items (SSI scores). These linear combinations were composed so that the correlation between both composed canonical variables was maximised. This correlation is called the canonical correlation (R_c). In other words, this investigated the research question to what extent the set of PCI scores be predicted or 'explained' by the pendant SSI scores. Transformation into z-scores was used to convert different raw scores of the items and scales in both batteries to share the same measurement unit with a mean of 0 and a standard deviation of 1 in order to make comparisons between sub-groups. Data analysis was performed using SPSS for Windows [34], LISREL [35] and SAS [36].

6.3 RESULTS

Sample

The source population consisted of patients who were referred by their general practitioner for a CAG in three hospitals in the northern part of the Netherlands. Of the 398 candidates screened for inclusion in this study, 139 (34.9%) did not return the first questionnaire. A questionnaire was received from the remaining 259 patients. We could not test the probability of systematic differences between non-responders and the study sample because no information was accessible without written informed consent from the patients who did not return the first questionnaire.

42 patients (16.2%) dropped out before the follow-up assessment. The reasons for not responding at follow-up were because the patient died (n= 7), had no heart failure (n=9), refused further participation (n=9), was too ill at follow-up (n=3), had moved (n=3) or did not react at all (n=11). To ensure that the patients who dropped out at follow-up did not deviate systematically from the study group, the characteristics of these patients at the time they returned the first questionnaire were

compared with the baseline characteristics of those who completed the questionnaire at follow-up. Except for educational level (the study sample had a statistically significant higher education), the demographic characteristics of the two groups were similar. This comparison also showed no statistically significant differences in mean scores on baseline health-status scales.
Analyses were based on 217 subjects (83.8%) who filled in the questionnaires at baseline and at post- test.

Demographics
The mean age of the patients was 60.6 (SD 9.43) with a range of 25 to 75. Sixty-one (28%) were female and 156 (72%) were male. Men were more likely to have a partner, to live together with someone, to have a higher education, and to be in employment. Five percent, 44%, 21% and 27%, respectively, had a self-reported NYHA-class I to IV at baseline. At follow-up, sixty-four (29%) had undergone a CABG, 71 (33%) a PTCA, and 82 (38%) were being treated with pharmacotherapy. Additional sample characteristics are presented in Chapter 5, table 5.1

Internal consistency
The homogeneity or unidimensionality of prospective change items and their corresponding transition items was estimated with the internal consistency coefficient (Cronbach's α). The multi-item transition scale 'physical functioning' yielded a somewhat higher internal consistency estimate (Cronbach's $\alpha = 0.92$) than the same scale composed of items' change-score (Cronbach's $\alpha = 0.86$), whereas these coefficients were identical for both versions of the scale 'emotional functioning' (Cronbach's $\alpha = 0.76$)

Convergent validity of prospective and retrospective measures of change
Two different operationalizations to capture the same concept of self-rated health were used. First, the classical repeated measurement change scores, derived from items from the MLHF-Q and the MOS-20. Second, a new set of operations to measure change in HRFS- domain by the means of transition items. We will consider measuring global self rated health with multi-item transition scales and hypothesized that, compaired to the serial change method, the underlying dimensions of physical and emotional functioning could be measured with the same results. One way to answer this question is to look at the concurrent or convergent validity of retrospective transition measures.
To investigate the convergent validity of the two domains assessed with repeated measurement as well with transition questions, we performed a factor comparison

by means of a principal component analysis with baseline items, item's difference scores and pendant transition items.
Table 6.2 presents the results of the principal component analysis (PCA) with items from three item-sets: 1. Items tapping the dimensions 'physical functioning' and 'emotional functioning' assessed at baseline, 2) the change score as the difference between baseline and follow-up for both dimensions (Serial Change Items) and 3) directly assessed change with the same items but modified to measure the perceived direction and magnitude of change at follow-up (Subjective Signified Items). Table 6.2 shows that in each battery, the items that cover the underlying construct of physical functioning have the expected factor loadings, [28,26] which were satisfactory >.50, except for the item with reference to the impact of heart failure on sleeping well at night. In each set of variables, the items that cover the domain of emotional functioning were, without exception, unambiguously related to the expected factor. The factor loadings of the prospective change scores on items were systematically lower than those estimated with the concordant baseline items and transition items, and the factor loadings of both baseline items and concordant transition items were of the same magnitude. Scales composed of multi-item change scores are usually more unreliable than either the baseline or follow-up measures on which they are based. The unreliability of the change score 'typically reduces its correlation with anything, including retrospective assessments' [37,38]. Notwithstanding these differences, the analysis of the three sets of items yielded similar factor loadings (Lambda in LISREL-notation) not only by this elaboration. The factor structure as prescribed by Rector et al. was confirmed by the baseline measure and the similarity between the factor structures of each mode of measurement series was shown by the overlap of the 95% confidence intervals.

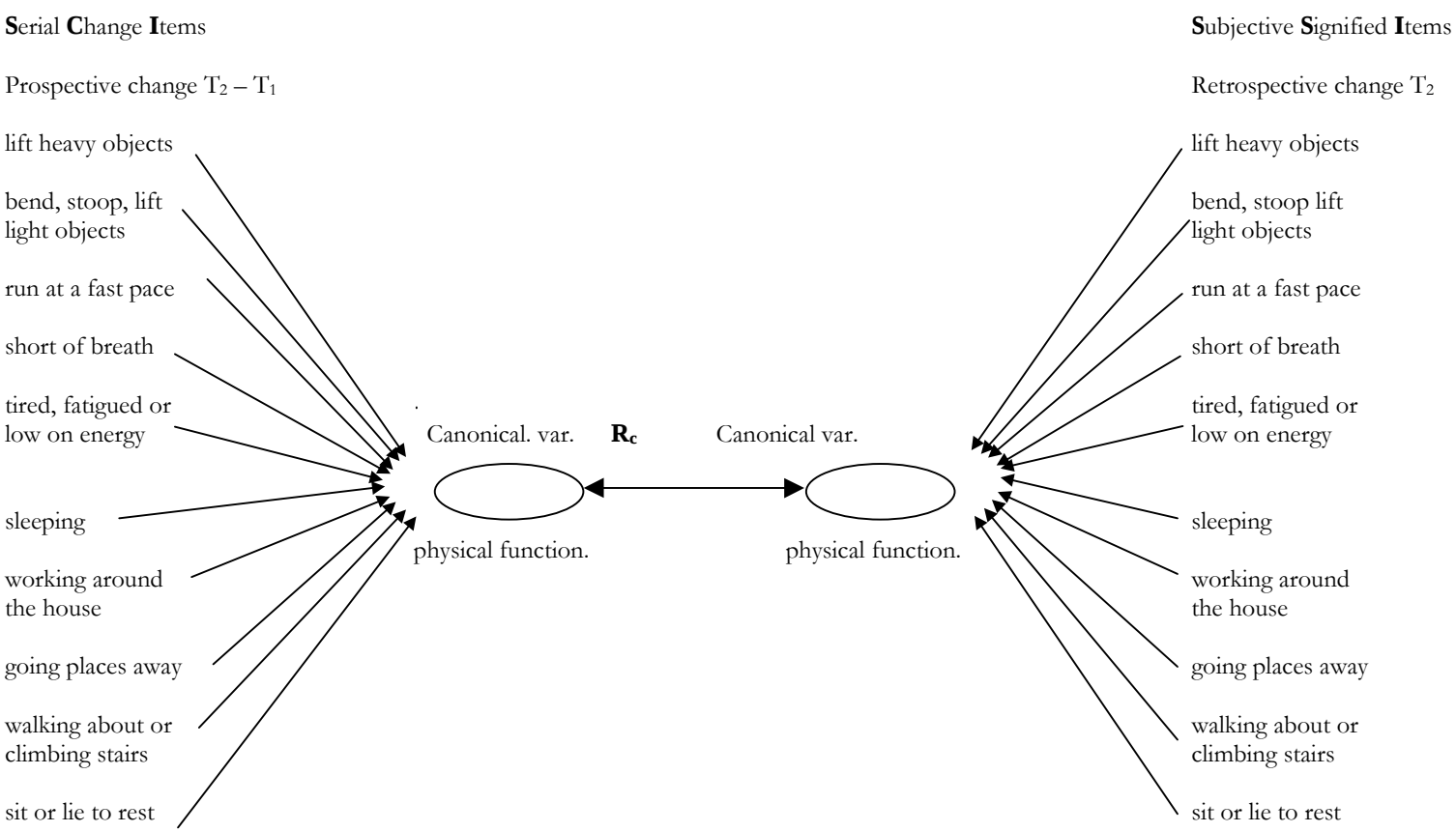
Table 6.2. *The loadings* of the scale components (item scale correlations) of baseline items, change scores of repeatedly assessed baseline items (Serial Change Items) and transition items (Subjective Signified Items) with 95% Confidence Intervals (C.I.)* (* Lambda in LISREL notation)

	Baseline Items			Serial Change Items (SCI)			Subjective Signified Items (SSI)		
	Loading	95%	C. I.	loading	95%	C.I.	loading	95%	C.I.
Physical functioning									
Sit or lie down to rest during the day	0.76	0,86	0,62	0,67	0,69	0,41	0,75	0,85	0,61
Walking about or climbing stairs	0.82	0,94	0,70	0,74	0,80	0,56	0,83	0,95	0,71
Working around the house or yard	0.86	0,98	0,74	0,71	0,76	0,52	0,79	0,89	0,69
Going places away from home	0.69	0,78	0,54	0.55	0,64	0,36	0,62	0,74	0,50
Sleeping well at night	0.54	0,63	0,35	0,39	0,47	0,19	0,49	0,56	0,28
Tired, fatigued or low on energy	0.74	0,83	0,59	0,71	0,76	0,52	0,68	0,80	0,56
Short of breath	0.80	0,92	0,68	0,70	0,82	0,58	0,81	0,93	0,69
Bend, stoop, or lift light objects	0.79	0,91	0,67	0,73	0,85	0,61	0,81	0,93	0,69
Lift heavy objects, like moving a table	0.79	0,91	0,67	0,75	0,84	0,60	0,78	0,90	0,66
Run at a fast pace	0.69	0,80	0,56	0,66	0,74	0,46	0,69	0,81	0,57
Emotional functioning MLHF-Q									
Feel you are a burden to the family	0,66	0,74	0,42	0,51	0,67	0,39	0,61	0,73	0,37
Feel a loss of self-control	0,85	0,97	0,73	0,88	1,00	0,76	0,84	0,98	0,70
Worry	0,72	0,84	0,60	0,69	0,83	0,55	0,64	0,78	0,50
Making it difficult for you to concentrate	0,60	0,72	0,41	0,59	0,70	0,35	0,57	0,65	0,37
Feel depressed	0,71	0,83	0,59	0,61	0,75	0,47	0,64	0,78	0,50

Canonical correlation between composites of change and transition items

Another way to answer the question 'do both operationalizations capture the same change in domains of health?' we also investigated with canonical correlation coefficients the magnitude of association between two concordant sets of items. Given that the transition items were forced into line with the content of the items used to assess prospective change we analysed both batteries of items with canonical correlation analysis (see Figure 6.2). Canonical coefficients only reflect the extent to which each item (given the other items) contributes to the composite of the items in the set of variables to which the item belongs. Therefore, we prefer to present the results with the canonical loadings as follows [39,40] : The correlation between the canonical variable and the items that belong to the underlying dimension of health status which were calculated with prospective change scores of scale items and corresponding transition items.

Figure 6.2: *Hypothetical model of canonical correlation analysis with serial and retrospective change-items of the physical functioning dimension.*



The canonical correlation (R_c) between the composite of the change score items and the composite of the concordant transition items that belong to the domain of physical functioning was R_c=.63 ($p < .001$), whereas the association between the canonical variables as the combination of the emotional functioning-items yielded a R_c=.48 ($p< .001$). These canonical correlations in terms of the percentage of linearly explained variance is fairly large between the two dimensions (40% and 23%, respectively).

Table 6. 3 *Correlations between the Serial Change-Items and the Subjective Signified Items and their corresponding canonical variable.*

	Serial Change Items (SCI)	Subjective Signified Items (SSI)
Physical functioning MLHF-Q		
	.55	.48
Sit or lie down to rest during the day		
Walking about or climbing stairs	.79	.90
Working around the house or yard	.68	.73
Going places away from home	.35	.34
Sleeping well at night	.36	.35
Tired, fatigued or low on energy	.55	.69
Short of breath	.41	.40
Bending, stooping, or lifting light objects	.55	.45
Lifting heavy objects, like moving a table	.77	.56
Running at a fast pace	.51	.44
Emotional functioning		
Feel you are a burden to the family	.48	.44
Feel a loss of self control	.64	.65
Worry	.61	.79
Making it difficult for you to concentrate or remember things	.60	.40
Feel depressed	.94	.94

Table 6.3 shows the correlations between the canonical variables with the original items: the canonical loadings. The prospective change items as well as the concordant transition items showed canonical loadings which were satisfactory according to the criterion > .30 of Levine [39] and, regardless the method of measuring change over time, we could unambiguously identify the items assessing the underlying dimensional configuration demonstrated with factor analysis. This result was remarkable because in the canonical correlation analysis the criterion of the linear combination is aimed at maximising the explained variance between two sets of variables, whereas in Principal Components Analysis (PCA), the linear combination criterion is aimed at maximising the explained variance within a set of variables. In spite of these divergent criteria underlying the method of data reduction, in the set with the change score items (PCI) as well as in the concordant set with transition items (SSI), identical items covered change in the expected domain of health. This result justifies, additional to the results from lisrel-analysis, the comparability of the summed composite of prospective change scales (PCS) with the summed composite of the retrospectively perceived change items (SSS).

Known groups validity

Another question was 'do both operationalizations of the measures have an equal ability to discriminate between subgroups known to differ on a clinically relevant variable?'. Therefore, improvement of angina pectoris (AP) was used as an external criterion to distinguish patients who improved from patient whose AP class stayed the same. Hence. the sample was divided into two groups who should differ based on the improvement of angina pectoris according to the NYHA classification [41]: patients who improved and patients who showed no shift in their NYHA classification. To test the hypothesis that both instruments of prospective change scales and retrospective transition scales have an equal ability to discriminate between these subgroups with known change in AP the Mann-Whitney U test was employed. For this analysis we have, apart from the overall scale of 10 all items belonging to the physical functioning dimension, used the MOS-20 items as an additional measure of physical functioning.

The results of the evaluation of the ability of the prospective and retrospective scales to discriminate between 'known groups' are reported in table 6.4 All p-values were beyond < .01, and effect sizes indicated small or moderate differences on both scales (small effect: ES < 0.20; moderate effect: ES > 0.20 < 0.50 and large effect: ES > 0.80). The difference in change in disease-specific physical functioning derived from repeated measurement detected a small difference between the groups whereas the generic scale estimated a moderate effect. Retrospective scales composed of the same

items showed effect size in reverse order. Change in the domain of the emotional functioning showed small difference between improved and stable AP groups for both methods.
This finding makes it reasonable to suppose that both methods are similar in several respects and that it is difficult to prove that one of them has superior qualities.

Table 6.4 *Discriminative ability of Serial Change Scales and corresponding retrospective transition scales between groups differing in change on the NYHA-classification (improved patients vs. patients who remained the same).*

	Improved Mean (SD)	N	stable mean (SD)	N	z-value	p-value	effect size[1]
Prospective							
Physical functioning:							
MLHF-Q scale	0.22 (1.01)	124	- 0.17 (0.97)	86	- 2.69	< 0.01	0.40
MOS-20 scale	0.27 (1.08)	119	- 0.24 (0.87)	86	- 3.42	< 0.01	0.53
Overall scale	0.25 (1.02)	124	- 0.20 (0.95)	86	- 2.93	< 0.01	0.46
Emotional functioning:							
MLHF-Q scale	0.21 (0.84)	123	- 0.19 (1.05)	87	- 2.42	< 0.01	0.41
Retrospective							
Physical functioning:							
MLHF-Q scale	0.39 (1.01)	123	- 0.27 (0.91)	86	- 4.74	< 0.01	0.71
MOS-20 scale	0.24 (1.11)	124	- 0.16 (0.88)	86	- 3.39	< 0.01	0.42
Overall scale	0.36 (1.03)	123	- 0.25 (0.91)	86	- 4.50	< 0.01	0.64
Emotional functioning:							
MLHF-Q scale	0.22 (1.03)	124	- 0.17 (0.96)	86	-.3.37	< 0.01	0.39

1 Effect size for independent samples: $(\bar{X}_1 - \bar{X}_1 / SD_{pooled})$ $SD = \sqrt{\frac{(S^2_1 + S^2_2)}{(N_1\text{-}1 + N_2\text{-}1)}}$

6.4 DISCUSSION

In patients suffering acute and or traumatic events, the measurement of health outcome due to clinical intervention is often hampered by the absence of a baseline measure. Reliable and valid measures of retrospective change would provide a solution to this problem. This study provides a method to assess perceived change with multi-item transition scales and we suggest that, in contrast with the single-item approach, it can provide a more valid determination of the change score that is signified as relevant by patients.

In most of the studies, transition questions are used as a single item to assess retrospective perceived change. One of the problems with the global assessment of change with one item is that the reliability cannot be established since Cronbach's alpha cannot be computed for a single item. Following Norman et al. [1], we could not locate studies that have examined the reliability of transition scales measuring a hypothesised underlying dimensional configuration. In our study, in which we applied multi-item measures of prospective and retrospective change in domains of HRFS, the scales had a satisfactory level of reliability [42]. The lower reliability of the emotional functioning scales may be due to a smaller number of items as the primary way to make scales more reliable is to make them longer [42].

Both the methods we applied in this study have several strengths and criticisms. Working with repeated measures and directly derived change scores has the major problem that they are ridden with a regression effect and prone to measurement error [43-45]. Change scores derived from repeated measurement may also be flawed by floor and ceiling effects [6,13], carryover effects of learning if the retest intervals are too short, specific events occurring between the first and second assessment, the 'natural' course of the disease, acquiescence and social desirability, and so on. [46]. Another threat to the validity of change scores is the assumption of researchers that subjects have an internalised perception of their level of functioning with regard to, for example, the domain of physical health status and that this internalised standard will not change from baseline to follow-up. This confounding is associated with response-shift bias [47]. There are also several threats to the reliability and validity of our findings based on the use of multi-transition items. First, retrospective perception of change may not be accurate due to recall bias. Patients may equate their present state with change in health status: if a respondent is doing poorly after treatment he might be inclined to think that on the whole things are getting worse even if his health state improved or did not change [1,48,49]. However, in the study of Fitzpatrick et al. [17], the transition questions were shown not to be determined by the patients' mood at follow-up or their present state. Additionally, patients who have suffered a large decline in health may overestimate their perception of baseline health

if they long for the time when their health was better [13,49]. Second, respondents may recalibrate their baseline situation due to the clinical intervention or may feel inclined to give socially desirable answers or they may change the "anchors" for their ratings over time, the so-called called 'response shift' [13,18]. Inaccurate recall seems to be determined by the time interval since exposure or intervention and by the degree of detail required [50], but the significance, the vividness and meaningfulness of events also contribute to recall. Some of the findings of Aseltine et al. [49] suggest that their measures assessing more concrete aspects of patient's condition provided greater correspondence between prospective and retrospective assessment than the more abstract measures of general health. Despite the limitations of transition questions, there is a growing realization that patients can be more directly involved in judging for themselves whether treatments have improved their health status or that relative to the observed health status of other patients by directly asked transition questions [5,6,17,51-53]. Moreover, transition questions were shown to be more sensitive to changes over time in health-related quality of life than were change scores [10,11,16]. The result of the analysis of equal factor loadings indicated that the multi-item transition indices (scales) measure phenomena similar to those measured by the serially assessed dimensions to which they were paired. This could arouse criticism because some items may not have contributed optimally to the assessment of underlying constructs. This may have been caused by the translation of items into transition questions, which may have changed their content, leading to the association of an item with the domain to which it did not belong. For example, the modification of the impact of heart failure on sleeping as an item belonging to the dimension of physical functioning may have become associated with sleeping problems caused by worries and anxiety. In several studies, the agreement between retrospective assessments and serial assessments were poor if single items were used. Therefore, the results of this study argue for multi-item batteries of transition items measuring (disease) specific and relevant domains of HRFS, since one-item transition questions do not cover the sum of aspects of health that belong to the underlying construct or dimension.
Further studies should address the psychometric aspects of transition scales used repeatedly in longitudinal studies, such as test-retest reliability and so on.
In conclusion, retrospectively assessed perception of change after intervention, appears to provide reliable and valid information compared to prospective change scores derived from repeated baseline measurement of health-status dimensions.

Reference List

1. Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. J Clin Epidemiol 1997; 50:869-879.
2. Beaton DE, Hogg-Johnson S, Bombardier C. Evaluating changes in health status: Reliability and responsiveness of five generic health status measures in workers with musculoskeletal disorders. J Clin Epidemiol 1996; 50:79-93.
3. Guyatt GH, Eagle DJ, Sackett B, Willan A, Griffith L, McIlroy W, et al. Measuring quality of life in the frail elderly. J.Clin.Epidemiol. 1993; 46:1433-1444.
4. 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. J Clin Epidemiol 1998; 51:617-631.
5. Redelmeier DA, Guyatt GH, Goldstein RS. Assessing the Minimal Important Difference in Symptoms: A Comparison of Two Techniques. Journal of Clinical Epidemiology 1996; 49:1215-1219.
6. Bindman AB, Keane D, Lurie N. Measuring health changes among severely ill patients; The floor phenomenon. Medical Care 1990; 28:1142-1152.
7. Stucki G, Liang MH, Fossel AH, Katz JN. Relative responsiveness of condition-specific and generic health status measures in degenerative lumbar spinal stenosis. J.Clin.Epidemiol. 1995; 48:1369-1378.
8. 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; 223-234.
9. Deyo RA, Inui TS. Toward Clinical Applications of Health Status Measures: Sensitivity of Scales to Clinically Important Changes. Health Services Research 1984; 19:275-289.
10. 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; 1202-1205.
11. Ziebland S. Measuring changes in health status. In: Jenkinson C, editor. Measuring health and medical outcomes. London: UCL Press, 1999:
12. MacKenzie RC, Charlson ME, DiGioia D, Kelley K. A patient-specific measure of change in maximal function. Arch Intern Med 1986; 146:1325-1329.
13. Baker DW, Hays RD, Brook RH. Understanding changes in health status;Is the floor phenomenon merely the last step of the staircase? Medical Care 1997; 35:1-15.
14. MacKenzie RC, Charlson ME, DiGioia D, Kelley K. Can the Sickness Impact Profile measure change? An example of scale assessment. J Chron Dis 1986; 39:429-438.
15. 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; 16:139-144.
16. Fischer D, Stewart AL, Bloch DA, Lorig K, Laurent D, Holman H. Capturing the patient's view of change as a clinical outcome measure. JAMA 1999; 282:1157-1163.

17. Fitzpatrick R, Ziebland S, Jenkinson C, Mowat A. Transition questions to assess outcome in rheumatoid arthritis. British Journal of Rheumatology 1993; 32:807-811.
18. Manusco CA, Charlson ME. Does recollection error threaten the validity of cross-sectional studies of effectiveness? Medical Care 1995; 33:AS77-AS88
19. Doll HA, Black NA, Flood AB, McPherson K. Criterion validation of the Nottingham Health Profile: Patient views of surgery for benign prostatic hypertrophy. Soc.Sci.Med. 1993; 37:115-122.
20. Cunny KA, Perri M. Single-item vs. multiple-item measures of health-related quality of life. Psychol Rep 1991; 69:127-130.
21. Kempen GIJM, Miedema I, van den Bos GAM, Ormel J. Relationship of domain-specific measures of health to perceived overall health among older subjects. J Clin Epidemiol 1998; 51:11-18.
22. Kempen GIJM. The MOS Short-Form General Health Survey: single item vs. multiple measures of health-related quality of life; some nuances. Psychol Rep 1992; 70:608-610.
23. Emery CF, Blumenthal JA. Perceived change among participants in an exercise program for older adults. The Gerontologist 1990; 30:516-521.
24. Guyatt GH. Measurement of health-related quality of life in heart failure. JACC 1993; 22:185A-191A.
25. Guyatt GH. Measurement of health-related quality of life in heart failure. Special Issue: Heart disease: The psychological challenge. The Irish Journal of Psychology 1994; 15:148-163.
26. 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 2000; 15: 380-391
27. Rector TS, Tschumperlin LK, Kubo SH, Bank AJ, Francis GS, McDonald KM, et al. Use of the Living With Heart Failure questionnaire to ascertain patients' perspectives on improvement in quality of life versus risk of drug-induced death. J.Card.Fail. 1995; 1:201-206.
28. Rector TS, Cohn JN. Assessment of patient outcome with the Minnesota Living with heart Failure questionnaire: Reliability and validity during a randomized, double blind, placebo-controlled trial of pimobendan. American Heart Journal 1992; October,124:1017-1025.
29. Stewart AL, Hays RD, Ware JE. The MOS Short-form General Health Survey: Reliability and validity in a patient population. Medical Care 1988; 26:724-735.
30. Noe LL, Vreeland MG, Pezzella SM, Trotter JP. A pharmacoeconomic assessment of Torsemide and Furosemide in the treatment of patients with congestive heart failure. Clinical Therapeutics 1999; 21:854-866.
31. Kubo SH, Gollub S, Bourge R, Rahko P, Cobb F, Jessup M, et al. Beneficial effects of Pimobendan on exercise tolerance and quality of life in patients with heart failure. Circulation 1992; 85:942-849.
32. Rector TS, Kubo SH, Cohn JN. Validity of the Minnesota Living with Heart Failure Questionnaire as a measure of therapeutic response to Enalapril or placebo. American Journal of Cardiology 1993; 71:1106-1107.

33. Rector TS. Effect of ACE inhibitors on the quality of life of patients with heart failure. Coron.Artery.Dis. 1995; 6:310-314.
34. Statistical Package for the Social Science. SPSS® for Windows,V7.5.3.Chicago:SPSS,inc. 1997;
35. LISREL 7® A guide to the program and applications.Chicago:Jöreskog and Sörbom SPSS inc. 1989;
36. SAS Institute,Inc.,SAS/STAT® User's Guide,Version 6,Fourth Edition,Volume 1,NC:SAS Institute Inc. 1990;
37. Cronbach LJ, Furby L. How we should measure "change"-or should we? Psychological Bulletin 1970; 74:68-80.
38. Bausell RB, Berman BM. Assessing patients' views of clinical changes [letter]. JAMA 2000; 283:1824
39. Levine MS. Canonical analysis and factor comparison. 1977; Beverly Hills: Sage Publications. 0-8039-0655-2.
40. Thompson B. Canonical correlation analysis: Uses and interpretation. 1984; Beverly Hills: Sage Publications. 0-8039-2392-9.
41. Criteria Committee of the New York Heart Association. Nomenclature and criteria for diagnosis of diseases of the heart and blood vessels: 1973; Boston: Little Brown.
42. Nunnally JC. Psychometric Theory. 2nd ed. New York: Mc Graw Hill, 1978.
43. Nunnally JC. The study of change in evaluation research: principles concerning measurement, experimental design, and analysis. In: Struening EL, Brewer MB, editors. Handbook of evaluation research. SAGE, 1983:231-269.
44. Gottman JM, Rushe RH. The Analysis of Change: Issues, Fallacies, and New Ideas. Journal of Consulting and Clinical Psychology 1993; 61:907-910.
45. Hsu LM. Regression Toward the Mean Associated With Measurement Error and the Identification of Improvement and Deterioration in Psychotherapy. Journal of Consulting and Clinical Psychology 1995; 63:141-144.
46. Cook TD, Campbell DT. Quasi-experimentation. Design & analysis issues for field settings. Chicago: Rand McNally College Publishing Company, 1979.
47. Schwarz CE, Sprangers MAG. Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. Soc.Sci.Med. 1999; 48:1531-1548.
48. Streiner DL, Norman GR. Health measurement scales. A practical guide to their development and use. second edition ed. Oxford: Oxford University Press, 1995.
49. Aseltine RH, Carlson KJ, Fowler FJ, Barry MJ. Comparing prospective and retrospective measures of treatment outcomes. Medical Care 1995; 33(uppl.):AS67-AS76
50. Coughlin SS. Recall bias in epidemiologic studies. J.Cinical Epidemiology 1990; 43:87-91.
51. Bjorner JB, Kristensen TS. Multi-item scales for measuring global self-rated health. Investigating of construct validity using structural equation models. Research On Aging 1999; 21:417-439.
52. Fitzpatrick R, Albrecht G. The plausibility of quality-of-life measures in different domains of health care. In: Nordenfelt L, editor. Concepts and measurements of quality of life in health care. Kluwer Academic Publishers, 1994:201-227.

53. Redemeier DA, Lorig K. Assessing the clinical importance of symptomatic improvements. An Illustration in Rheumatology. Arch Intern Med 1993; 153:1337-1342.