Quality of Life Research (2019) 28:2609–2620
https://doi.org/10.1007/s11136-019-02195-9

Response shift effects in patients' assessments of their quality of life after cardiac rehabilitation

Michael Friedrich[1] · **Jan Karoff**[2] · **Andreas Hinz**[1]

Accepted: 30 April 2019 / Published online: 8 May 2019
© Springer Nature Switzerland AG 2019

Abstract

Purpose The effect of intervention programs on health-related quality of life (HRQoL) can be underestimated due to response shift effects. This study aims to compare HRQoL between cardiac patients taking part in a rehabilitation program and the general population and to investigate changes in HRQoL in terms of response shift with two approaches.

Methods A sample of 282 cardiac rehabilitation inpatients (response rate: 58.9%) responded to the self-report quality of life questionnaire EORTC QLQ-C30 at baseline (during rehabilitation) and three months later (actual and retrospective judgment). Their HRQoL was compared to that of the general population. Response shift evaluation complemented the thentest with the structural equation modeling approach.

Results Compared to the general population, patients showed impaired quality of life on all scales (Hedges' g between 0.31 and 1.57). The complementation of the thentest with the structural equation modeling approach revealed response shift effects in physical, emotional, cognitive, and social functioning. No effects were found in role functioning.

Conclusions The combination of both the thentest and the structural equation modeling approaches proved to be essential for obtaining comprehensive statistical evidence that response shift can distort measurements of change. Our results suggest that studies that use the thentest to evaluate the effectiveness of interventions should complement their analyses with the structural equation modeling approach to avoid biased effects.

Keywords Quality of life · Response shift · Thentest · Structural equation modeling

Background

A central goal of cardiovascular care is to improve patients' health status. In addition to mortality and morbidity outcomes, patient-reported health-related quality of life (HRQoL) is an important measure of health, especially when examining the effects of interventions on cardiovascular health [1]. HRQoL can predict mortality [2], cardiovascular events, hospitalization, and costs of care [1]. Patients with a cardiovascular disease (e.g., congenital heart disease [3], congestive heart failure [4], myocardial infarction [5], or coronary heart disease [6]) have impaired HRQoL compared to the general population.

One common intervention in cardiovascular care is cardiac rehabilitation [7]. In Germany, it is covered by public and private health insurers as well as the German Statutory Pension Insurance Scheme [8]. Methods of treatment comprise preventive cure (Heilverfahren) and—more frequently—follow-up treatment immediately after acute cardiac events (Anschlussheilbehandlung). Though outpatient rehabilitation services were introduced in the 2000s, it is far more common for patients to receive inpatient services for a period of time typically lasting three or more weeks.

The effect of an intervention aimed at improving HRQoL is usually quantified as the difference between measurements taken at baseline and follow-up assessments. The most common approach only takes the actual measurements of perceived HRQoL into account (e.g., the mean difference *posttest–minus–pretest*) and provides a more objective

✉ Michael Friedrich
michael.friedrich@medizin.uni-leipzig.de

Jan Karoff
karoff@uni-wuppertal.de

Andreas Hinz
andreas.hinz@medizin.uni-leipzig.de

[1] Department of Medical Psychology and Medical Sociology, University of Leipzig, Leipzig, Germany

[2] Institute of Educational Sciences, University of Wuppertal, Wuppertal, Germany

measure, one of *observed change*. Another approach considers the fact that a person's perception of the quality in question can change over time, even if the quality itself does not change. Hence, this approach does not use measurements of perceived HRQoL at baseline. Instead, it uses a thentest, a retrospective assessment of baseline HRQoL that is first reported at follow-up (e.g., [9–11]). The difference (*posttest–minus–thentest*) provides a more subjective measure, one of *perceived change*, which is more meaningful for understanding the effects of interventions, as patients perceive them. Each of these three measurements (pretest, posttest, thentest) is based on its own frame of reference, and every mean difference can be biased. This bias is differentiated into four types (reconceptualization, reprioritization, and uniform and non-uniform recalibration) and is called response shift [12]. Reconceptualization describes a redefinition of the target construct, reprioritization describes a change in the importance of the target components, and recalibration describes a change in the internal standards. Recalibration is called uniform if the change in internal standards can be explained by change in the target construct, and it is called non-uniform if not. One method for analyzing these types of bias is the structural equation modelling (SEM) approach introduced by Oort [13]. Theoretically, response shift can be caused by different mechanisms such as, among others, coping and social comparison [14]. Since it is an aim of cardiac rehabilitation to support active coping, response shift should be expected to occur in cardiac rehabilitation [7]. Dempster et al. [7] showed that response shift does indeed occur during cardiac rehabilitation and concluded that this bias probably leads to an underestimation of the effects of the intervention. Because Schwartz et al. [15] pointed out that the thentest is susceptible to recall bias and potentially contaminated by other influences (e.g., social desirability, effort justification, and implicit theories of change), the question arises how a person's recollection differs between pretest and thentest besides response shift. This question can be answered when complementing both approaches (investigating observed and perceived change) into one approach using SEM. The complementary integration of these two approaches is new, because so far either only one kind of change has been examined, or both kinds of change have been compared as competing methods, e.g., Visser et al. [16]. Considering the two approaches as complementary within a single structural equation model and not as competing approaches has the advantage that the susceptibility to memory distortion of the thentest approach can be quantified.

In light of these theoretical implications and empirical findings, our aims are as follows: (a) to examine differences in HRQoL between patients undergoing cardiac rehabilitation and the general population, (b) to investigate changes in HRQoL that were observed and that were perceived, and (c) to explore response shift effects and indications of recall bias.

Methods

Study participants

Between February 2015 and April 2016, a group of 479 cardiac rehabilitation inpatients treated in a German rehabilitation clinic administrated by the Deutsche Rentenversicherung Westfalen were asked to participate in the study. Inclusion criteria were (1) survival of an acute cardiovascular event, (2) age of 18 years or older, and (3) the absence of any severe cognitive or verbal impairments that would interfere with a patient's ability to complete questionnaires. Informed consent was obtained from the study participants after they were given an explanation of the purpose of the study and data collection and storage methods. Of the 479 patients invited to take part in the study, 356 (74%) consented to participate and filled in the first questionnaire (baseline). Three months later, these patients were sent a packet by mail including a letter, a questionnaire (follow-up), and a stamped addressed return envelope. If they did not respond, they were sent one reminder by mail. In total, data from 282 patients (79%) were available for analysis. The study was approved by the Ethics Committee of the University of Leipzig.

General population

The reference data were taken from two studies that examined representative samples of the German general population [17, 18] ($n = 4476$). From these, a subsample was selected (1760 males, 343 females) so that the proportion of general population females was identical with that of our rehabilitation patients' baseline sample (16.3%) and that the mean age of the general population sample was very similar to that of the patients' baseline sample ($M = 55.6$ years). The selection was realized by systematically excluding young participants and women from the original general population sample until the distribution of the patients' sample was reached.

Instruments

The sociodemographic characteristics we accounted for included: gender, age at baseline, education, employment status, and partnership status. The medical characteristics we recorded were diagnosis and time since start of treatment (in weeks).

HRQoL was measured with the functioning scales of the Quality of Life Questionnaire EORTC QLQ-C30 that was

developed by the European Organization for Research and Treatment of Cancer (EORTC). Although this is a disease-specific instrument developed for use with cancer patients, it can also be used to assess HRQoL in other populations as well including the general population [17, 19–22] and other patient groups suffering from, for example, chronic pain [23] or cardiac diseases [24]. The instrument contains 30 items distributed across five functioning scales, three symptom scales, six symptom items, and one global health/quality of life scale [25]. All scores are linearly transformed to obtain the range 0–100. Higher values on the functioning scales indicate higher functioning, and higher values on the symptom scales indicate greater levels of burden [26]. A recent study tested the higher order measurement structure [27].

Statistical analyses

Missing values were estimated using the Expectation Maximization procedure [28]. Statistical analyses were performed using IBM SPSS Statistics 23, IBM SPSS Amos 23, using the maximum likelihood estimation procedure, and Microsoft EXCEL 2010 supplemented by the "Real Statistics Resource Pack" for EXCEL [29].

Comparisons of means were conducted with *t* tests for independent groups (general population) and the respective *t* tests for dependent groups (between pretest, posttest and thentest). Furthermore, we computed effect sizes (Hedges' *g*) to express the mean score differences in relation to the pooled standard deviation. Hedges' *g* is a bias-corrected value of Cohen's *d* and is classified with $g \geq 0.2$ as small, $g \geq 0.5$ medium, and ≥ 0.8 large [30]. Type-I-error probabilities (*p* values) for the effect sizes were computed using their standard errors [31].

Detection of response shift was conducted with the SEM approach proposed by Oort [13, 32]. First, the measurement model of functioning quality of life according to Gerlich et al. [33] was tested for each measurement (pretest, posttest, and thentest) separately. This model included the respective five EORTC QLQ-C30 functioning scales (physical functioning, role functioning, social functioning, cognitive functioning, and emotional functioning). Then, these three models were combined through introducing between occasion covariances for each scale and additional within occasion covariances between the residuals of physical and role functioning, and between role and social functioning, analogous to Gerlich et al. [33]. The model diagram used for response shift evaluation is presented in Fig. 1. Subsequently the response shift detection process is based on the following three steps that are distinguished through models containing different levels of restriction:

(1) *Unconstrained model* Here the latent variables of all three measurements are fixed to a mean of 0 and a variance of 1 to fully identify the model. This model serves as a baseline model for comparisons with the fully constrained model (next step). If the unconstrained model does not show acceptable fit, reconceptualization response shift between measurements is indicated, and the analysis ends here, because it is not possible to assume comparable concepts in general.

(2) *Fully constrained model* This model assumes the null hypothesis (no response shift). Accordingly all parameters (weights, intercepts and residual variances) are constrained to be equal across all three measurements. Here, only the latent variable of the pretest measurement is fixed to a mean of 0 and a variance of 1 to identify the model. Acceptable fit indicates no further types of response shift, and the analysis ends at this step. In the case of poor fit, the following step is taken.

(3) *Response shift model* In this model, restrictions from the previous model are freed one after another. A restriction identified as misspecified was released when the release led to a substantial improvement of the model fit. The sequence of releasing begins with residual variances (to detect non-uniform recalibration), followed by the intercepts (uniform recalibration), and then by the weights (reprioritization). The releasing is done until there is no substantial increase in model fit. A released parameter is indicative of response shift, and the type of the released parameter determines the type of response shift (residuals: *non-uniform recalibration*, intercepts: *uniform recalibration*, weights: *reprioritization*).

Unacceptable misspecifications were identified using the combination of the modification index, the power of the MI-Test, and the expected parameter change [34]. According to Saris et al.'s suggestion [34], we chose the following critical deviations: ten percent of the pretest sample's variance of the respective functioning scale for *residual variances*; ten percent of the pretest standard deviation of the scale for *regression weights*; and for *intercepts*, we followed the guidelines for interpreting the EORTC QLQ-C30 change scores that were proposed by Cocks et al. [35]. Type-I-error-probability was set to 0.05, and high power to 0.80. Model fit was assessed with a combinational rule of CFI (comparative fit index) and SRMR (standardized root–mean-square residual) [36]. Models were rejected if both CFI and SRMR indicate poor fit, that is, if $CFI < 0.95$ and $SRMR > 0.08$. To indicate the trade-off between model fit and model complexity, we additionally present AIC (Akaike information criterion). To evaluate model differences, a value of $\Delta CFI \geq 0.002$ was regarded as substantial model improvement [37].

To judge the share of response shift within the change that obviously occurred between two means, it is helpful to decompose the change into two parts, one part that indicates the

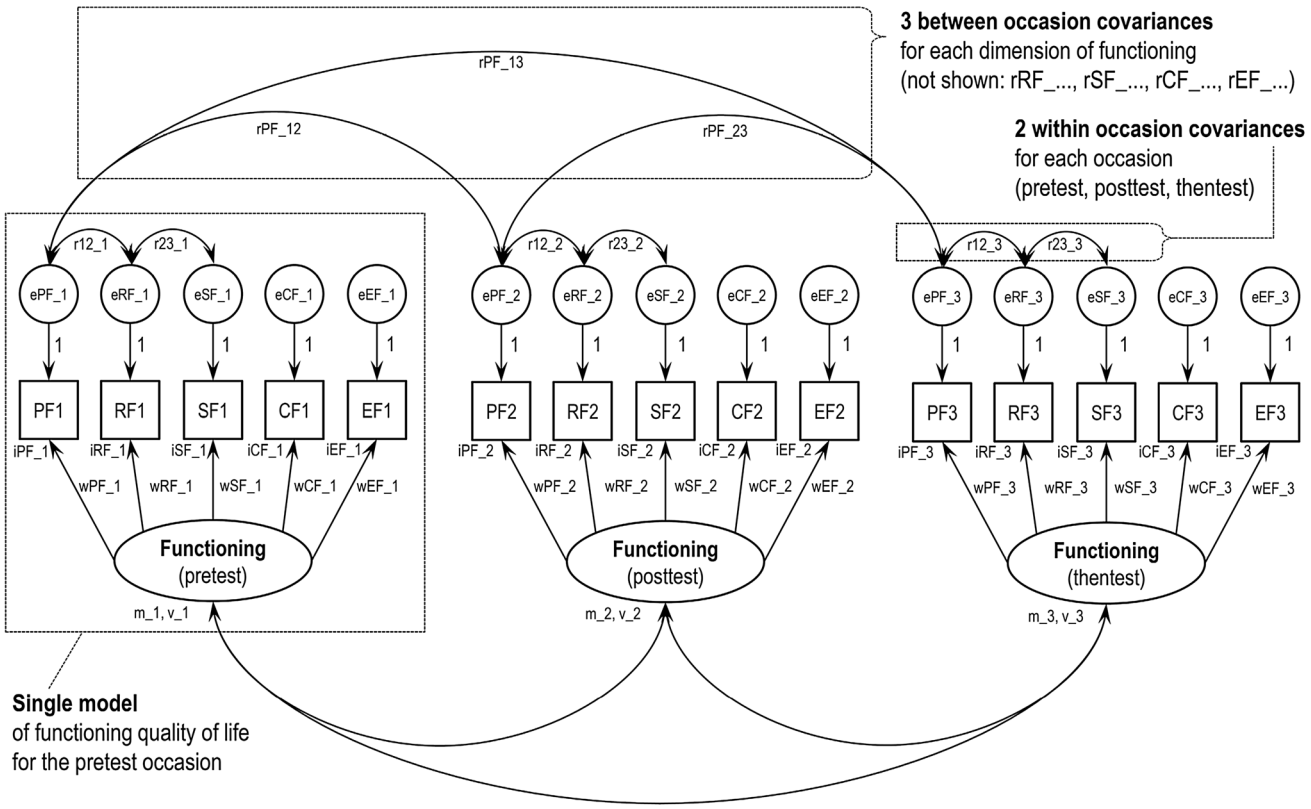

Fig. 1 Diagram of the model for response shift evaluation. Abbreviations: *PF* physical functioning, *RF* role functioning, *SF* social functioning, *CF* cognitive functioning, EF emotional functioning. Annotations: *Rectangles* manifest variables, *ovals* latent variables, *circles* residuals, *straight arrows* regression weights, *curved arrows* covariances. Terminology of the model parameters: *r* covariance, *e* residual variance, *i* intercept, *w* regression weight, *m* latent mean, *v* latent variance, numbers after the underscore indicate the occasion: *1* pretest, *2* posttest, *3* thentest

difference under the assumption that no response shift would have occurred, that is, if the parameters would have been equal to the baseline measurement (called "true change"), and a another part that indicates the amount of response shift. The meaning of response shift for the mean of an item can be illustrated by the idea behind the SEM approach. Here a latent variable is assumed to manifest itself in a number of items. As consequence, the mean of every item can be decomposed into three components: a common, a unique, and a residual quality. However, only the first two are of relevance, because they are shares of response shift that affect the item's mean.

Oort [13] showed how to decompose the mean difference into three components: the contribution of true change, of uniform recalibration, of reconceptualization, and of reprioritization. Note that Oort [13] used the term "observed change" to indicate the change that is not decomposed yet and comprises true change and response shift. Because we use the term "observed change" to indicate the change between the actual baseline (pretest) and follow-up (posttest) measurements, we use the term "observed *difference*" to indicate the change that is not yet decomposed. Nevertheless, it is also an observed change. The following equation describes the decomposition of response shift effects:

$$
\begin{aligned}
&X_2 - X_1 \text{ Observed difference} \\
&= \left(i_2 - i_1\right) \text{ Uniform recalibration} \\
&\quad + \left(w_2 - w_1\right) \cdot L_1 \text{ Reprioritization/} \\
&\qquad \text{reconceptualization without change in } \mathrm{L}_2 \\
&\quad + \left(w_2 - w_1\right) \cdot \left(L_2 - L_1\right) \text{ Reprioritizationreconceptualization} \\
&\qquad \text{with change in } \mathrm{L}_2 \\
&\quad + w_1 \cdot \left(L_2 - L_1\right) \text{ True change}
\end{aligned}
$$

X denotes the mean of the observed score of a manifest variable. It can be decomposed to $X = i + w \cdot L + e$. The letter $\boldsymbol{i}$ denotes the intercept (constant), w the regression weight (factor loading, constant) between the latent variable L (factor score, mean) and the manifest variable X, and e indicates the residual factor score (mean), which is set to zero in the equation above, because the residual means are fixed to zero in the model. If the mean of L_1 is zero,

this equation reduces to that presented by Oort [13, Eq. 8, p.594]: $X_2 - X_1 = (i_2 - i_1) + (w_2 - w_1) \cdot (L_2) + w_1 \cdot (L_2)$.

Results

Sociodemographic and medical characteristics

Of the 356 baseline participants 282 (79%) returned the 3-month follow-up questionnaire. Data from the 74 (21%) participants who dropped out were excluded from the analyses. Table 1 presents the sociodemographic and medical characteristics of both samples (dropout and analysis sample). Column proportions between angiopathy patients who dropped out and those who completed the study differed to a statistically significant extent ($p < 0.05$): 16% in the dropout sample and 7% in the analysis sample. The majority of patients was male (83%), between 50 and 70 years old (73%), had 8 to 10 years of education (71%), and was employed (78%). The most common diagnosis was coronary heart disease (69%). The mean age of the study participants was 56.4 (SD = 8.2) years.

Comparison with the general population

Cardiac patients showed worse HRQoL than the general population in all dimensions. Table 2 presents mean scores, standard deviations, and effect sizes (Hedges' g) for both groups. Hedges' g for the functioning scales showed only large effects ($|g| \geq 0.80$) and ranged from -0.91 (physical functioning) to -1.57 (social functioning). Regarding the symptoms, Hedges' g showed higher levels of burden and ranged from 0.31 (nausea/vomiting) to 1.66 (dyspnoea). Besides dyspnoea, three further symptoms showed large standardized mean differences: fatigue ($g = 1.38$), financial difficulties ($g = 1.25$), and insomnia ($g = 0.99$). The global health/quality of life scale (QL) showed an effect of $g = -0.76$.

Observed change in HRQoL (posttest–minus–pretest)

Three months after cardiac rehabilitation the means of all five functioning scales were higher than the means that were reported during cardiac rehabilitation (Table 2). The effect sizes of the differences of the functioning scales from pretest to posttest were all positive and statistically significant ($p < 0.001$). Hedges' g ranged between 0.18 (cognitive functioning) and 0.29 (social functioning). This increase in HRQoL was accompanied by statistically significant declines in the symptoms dyspnoea ($g = -0.45$), fatigue ($g = -0.37$), pain ($g = -0.28$), and appetite loss ($g = -0.20$).

Perceived change in HRQoL (posttest–minus–thentest)

Three months after cardiac rehabilitation the perceived change in physical functioning was greater than the observed change ($g = 0.53$ perceived vs. $g = 0.24$ observed, Table 2). All other functioning scales showed lower perceived than observed change. Hedges' g for these scales ranged from 0.12 (cognitive functioning) to 0.23 (social functioning).

Detection of response shift effects

Using the thentest approach, we found response shift in the physical functioning domain with an effect of $g = 0.32$ (Table 2, column "pre–then"). To get a more comprehensive picture of the response shift effects, we also used the SEM approach complemented by the thentest. Consequently, we first analyzed the fit of the measurement model for each measurement (pretest, posttest and thentest) separately. The fit of the single models was acceptable: all measurements showed values of CFI > 0.96 and SRMR < 0.04 and therefore no indication of reconceptualization (Table 3). The unconstrained combined model confirmed this assumption with a slightly lower but also acceptable fit (CFI = 0.96 and SRMR = 0.07). The fit of the fully constrained model was marginally acceptable (CFI = 0.90 and SRMR = 0.07), but the decrease of fit (ΔCFI = -0.052) was substantial, indicating other types of response shift. After the release of six constraints, of which each led to a substantial increase in model fit, the final response shift model was found.

The step-by-step procedure revealed all remaining kinds of response shift. The resulting parameters of the response shift model are shown in Table 4. Reprioritization, a change in the importance of an item relative to the others [13], was found in cognitive and emotional functioning. The weights were higher at the pretest measurement. Uniform recalibration, a change in the respondent's internal standards of measurement [13], was indicated in physical and cognitive functioning. The intercept of physical functioning was lower in the thentest measurement, and the intercept of cognitive functioning was lower in the posttest measurement. With the fact that the intercepts of physical and cognitive functioning were different in the follow-up assessments, the question arises of why only one follow-up assessment was affected. It is conceivable that this might be a methodological artefact of restrictions, and if we had freed the intercepts of physical functioning (in pretest and posttest) and the intercepts from cognitive functioning (in pretest and thentest), the uniform recalibration would be distributed across both follow-up assessments. But this hypothesis did not hold. When we freed these parameters, the intercepts did change minimally (less than one raw point) and the model fit did not increase substantially.

Table 1 Sociodemographic and medical characteristics

	Dropout sample ($n=74$)	Analysis sample ($n=282$)
	N (%)	*N* (%)
Sociodemographic characteristics		
Gender	74 (100.0)	282 (100.0)
Male	65 (87.8)	233 (82.6)
Female	9 (12.2)	49 (17.4)
Age	74 (100.0)	282 (100.0)
25 to < 50 years	21 (28.4)	59 (20.9)
50 to < 60 years	38 (51.4)	146 (51.8)
60 to < 70 years	15 (20.3)	60 (21.3)
70 years and older	0 (0.0)	17 (6.0)
M (SD)	53.38 (8.19)	56.18 (8.02)
Education[a]	74 (100.0)	281 (100.0)
Elementary school (8–9 years)	39 (52.7)	113 (40.2)
Junior high school (10 years)	21 (28.4)	88 (31.3)
High school/university (> 10 years)	12 (16.2)	71 (25.3)
Other and no formal qualification	2 (2.7)	9 (3.2)
Employment status[a]	73 (100.0)	280 (100.0)
Employed	57 (78.1)	218 (77.9)
Unemployed	9 (12.3)	21 (7.5)
Retired	4 (5.5)	33 (11.8)
Other	3 (4.1)	8 (2.9)
Partnership[a]	73 (100.0)	279 (100.0)
Yes	56 (76.7)	224 (80.3)
Medical characteristics		
Time since start of treatment in weeks[a]	71 (100.0)	273 (100.0)
Up to 6 weeks	17 (23.9)	50 (18.3)
> 6 to 12 weeks	34 (47.9)	115 (42.1)
> 12 weeks	20 (28.2)	108 (39.6)
Median (IQR)	8.29 (13.14)	9.00 (35.43)
Diagnosis	74 (100.0)	282 (100.0)
CHD—coronary heart disease *with* infarction[1]	21 (28.4)	98 (34.8)
CHD—coronary heart disease *without* infarction[2]	28 (37.8)	95 (33.7)
Structural heart diseases[3]	6 (8.1)	32 (11.3)
Angiopathy[4]	12 (16.2)	19 (6.7)
Other diagnoses[5]	7 (9.5)	38 (13.5)
Cardiac surgery within the last 3 months[a]	74 (100.0)	280 (100.0)
Yes	16 (21.6)	87 (31.1)
Cardiac infarction within the last 3 months[a]	74 (100.0)	281 (100.0)
Yes	45 (60.8)	168 (59.8)

M Mean, *SD* standard deviation, *IQR* interquartile range

[a]Missing values considered

[1]ICD-10: I21-I23 and I25.2

[2]ICD-10: I24-I25 except I25.2

[3]Atherosclerosis, heart valve diseases, cardiomyopathy, unstable angina pectoris

[4]Aneurysm, pulmonary hypertension, embolic disease, thrombosis, stenosis

[5]Essential hypertension, stroke, arrhythmia, endocarditis, complications and others

Table 2 Mean scores and effect sizes

	M (SD)				Hedges'*g* (*p* value)			
	General population[a]	Pre[b]	Post[b]	Then[b]	Pre-GP	Post–pre[c]	Post–then[d]	Pre–then[e]
Functioning scales								
PF	90.11 (17.05)	*74.58 (17.71)*	*79.07 (19.28)*	*68.35 (20.85)*	**−0.91 (<0.001)**	**0.24 (<0.001)**	**0.53 (<0.001)**	**0.32 (<0.001)**
RF	88.10 (22.81)	*55.88 (29.30)*	*64.14 (32.48)*	*57.81 (32.06)*	**−1.36 (<0.001)**	**0.27 (<0.001)**	**0.20 (0.001)**	−0.06 (0.360)
SF	91.51 (19.23)	*59.04 (29.64)*	*67.51 (28.76)*	*61.02 (28.26)*	**−1.57 (<0.001)**	**0.29 (<0.001)**	**0.23 (<0.001)**	−0.07 (0.275)
CF	91.81 (16.44)	*70.85 (27.50)*	*75.71 (27.27)*	*72.59 (26.86)*	**−1.16 (<0.001)**	**0.18 (<0.001)**	**0.12 (0.003)**	−0.06 (0.197)
EF	81.28 (20.25)	*55.12 (29.70)*	*63.45 (28.60)*	*57.07 (28.12)*	**−1.21 (<0.001)**	**0.28 (<0.001)**	**0.22 (<0.001)**	−0.07 (0.169)
Symptom scales								
FA	16.52 (22.27)	47.97 (25.79)	38.55 (25.37)	44.72 (25.44)	**1.38 (<0.001)**	**−0.37 (<0.001)**	**−0.24 (<0.001)**	**0.13 (0.028)**
NV	2.22 (8.22)	4.92 (11.00)	4.98 (12.21)	4.61 (10.82)	**0.31 (<0.001)**	0.00 (0.949)	0.03 (0.547)	0.03 (0.698)
PA	17.49 (24.99)	36.83 (32.08)	28.22 (29.99)	37.20 (30.08)	**0.75 (<0.001)**	**−0.28 (<0.001)**	**−0.30 (<0.001)**	−0.01 (0.833)
DY	9.30 (21.58)	47.64 (32.13)	33.02 (32.69)	43.62 (34.91)	**1.66 (<0.001)**	**−0.45 (<0.001)**	**−0.31 (<0.001)**	0.12 (0.052)
IN	16.26 (26.65)	44.05 (36.19)	39.99 (37.05)	38.18 (34.27)	**0.99 (<0.001)**	−0.11 (0.058)	0.05 (0.278)	**0.17 (0.001)**
AP	4.70 (14.88)	18.09 (27.98)	12.97 (23.76)	14.89 (24.65)	**0.79 (<0.001)**	**−0.20 (0.006)**	−0.08 (0.168)	0.12 (0.067)
CO	2.89 (12.48)	8.44 (19.79)	7.61 (17.94)	7.14 (17.80)	**0.41 (<0.001)**	−0.04 (0.532)	0.03 (0.654)	0.07 (0.322)
DI	2.46 (10.73)	9.00 (19.83)	10.20 (20.11)	8.27 (18.29)	**0.54 (<0.001)**	0.06 (0.402)	**0.10 (0.044)**	0.04 (0.589)
FI	6.12 (18.59)	32.34 (33.82)	35.58 (35.78)	35.59 (34.42)	**1.25 (<0.001)**	0.09 (0.069)	0.00 (0.996)	−0.09 (0.071)
Summary scores								
QL	70.87 (21.52)	54.62 (20.78)	65.36 (22.35)	58.91 (19.98)	**−0.76 (<0.001)**	**0.50 (<0.001)**	**0.30 (<0.001)**	**−0.21 (0.001)**
SS		69.09 (16.73)	75.30 (18.15)	70.76 (16.99)		**0.35 (<0.001)**	**0.26 (<0.001)**	**−0.10 (0.046)**

[a] $n = 2103$, except with missing values in **NV** and **DI** with $n = 1$, **CO** with $n = 2$, **AP** with $n = 4$ and **FI** with $n = 5$; no data for **SS** in GP available

[b] $n = 282$; for functioning scales (italic): means and standard deviations that were implied by the response shift model

[c] Observed change (posttest–minus–pretest)

[d] Perceived change (posttest–minus–thentest)

[e] Difference of recollection (pretest–minus–thentest)

M (SD) Mean (standard deviation), *Hedges' g* effect size, bias corrected version of Cohen's d, *p value* type-I-error probability: significant values ($p < 0.05$) bold, *GP* general population, *pre* measurement at baseline (pretest), *post* measurement at follow-up (posttest), *then* retrospective measurement of pretest-QoL at follow-up (thentest); functioning scales: *PF* physical functioning, *RF* role functioning, *SF* social functioning, *CF* cognitive functioning, *EF* emotional functioning; symptom scales: *FA* fatigue, *NV* nausea/vomiting, *PA* pain, *DY* dyspnoea, *IN* insomnia, *AP* appetite loss, *CO* constipation, *DI* diarrhea, *FI* financial difficulties; Summary scores: *QL* quality-of-life scale (2 Items), *SS* summary score

Non-uniform recalibration was found in the physical and social functioning domains. While the residual variance of physical functioning was higher in the thentest measurement, the residual variance of social functioning was higher in the pretest measurement. The means of the latent variables (overall functioning quality of life) of the follow-up measurements changed from 0.00 (pretest, fixed) to 0.11 (thentest) and finally to 0.47 (posttest).

Decomposition of response shift effects and recall bias

Table 5 shows the decomposition of the observed differences into true change and contributions of response shift (uniform recalibration and reconceptualization/reprioritization). We present raw differences as well as the pooled standard deviations.

Response shift effects that influenced the observed difference were found in physical, cognitive, and emotional functioning. Regarding physical functioning, we found the influence of the uniform recalibration in the perceived change (lower intercept in thentest measurement) with + 7.27 points and in the difference of recollection (thentest–minus–pretest) with the opposite sign. This effect increased the perception of change and changed the direction of the systematic deviation between pretest and its retrospective pendant. Regarding cognitive functioning, two effects influenced the comparisons. The uniform recalibration (lower intercept in the posttest measurement) reduced the observed and the perceived change by − 2.60 points. The reprioritization effect (higher weight in the pretest measurement) lowered the observed change by − 2.20 points and the systematic deviation between the thentest and the pretest by − 0.51 points. This effect is different for both comparisons because it depends on the

Table 3 Response shift detection ($n = 282$)

	χ^2 (df)	p	χ^2/df	CFI	ΔCFI	SRMR	AIC
Single models							
M1) pretest	16.9 (3)	0.0008	5.6	0.9697	–	0.0397	50.9
M2) posttest	20.7 (3)	0.0001	6.9	0.9771	–	0.0355	54.7
M3) thentest	7.5 (3)	0.0588	2.5	0.9925	–	0.0250	41.5
Combined models[a]							
M4) Unconstrained	191.7 (66)	< 0.0001	2.9	0.9568	–	0.0662	329.7
(M5) Fully constrained	370.5 (92)	< 0.0001	4.0	0.9044	− 0.0524	0.0727	456.5
(M6) eSF_pretest	350.9 (91)	< 0.0001	3.9	0.9108	0.0064	0.0679	438.9
(M7) ePF_thentest	335.2 (90)	< 0.0001	3.7	0.9158	0.0050	0.0684	425.2
(M8) iPF_thentest	257.4 (89)	< 0.0001	2.9	0.9422	0.0264	0.0668	349.4
(M9) iCF_posttest	249.7 (88)	< 0.0001	2.8	0.9445	0.0023	0.0675	343.7
(M10) wEF_pretest	242.1 (87)	< 0.0001	2.8	0.9468	0.0023	0.0658	338.1
(M11) wCF_pretest[b]	232.3 (86)	<0.0001	2.7	0.9498	0.0030	0.0654	330.3

[a]Models M6 to M11 that are mentioned after the fully constrained model are nested. Every model contains one more parameter that is not restricted to be equal. Names of these models show the freed parameter: ***e*** residual variance, ***i*** intercept, ***w*** regression weight

[b]Response shift model (M11): non-uniform recalibration (residual variances of SF (pretest) and PF (thentest) free), uniform recalibration (intercepts of PF (thentest) and CF (posttest) free) and reprioritization (regression weights of EF (pretest) and CF (pretest) free)

χ^2 Chi squared-statistic (minimum discrepancy function), *df* degrees of freedom, *p* type-I-error-probability, *CFI* comparative fit index, Δ*CFI* fit-difference between successive models, *SRMR* standardized root–mean-square residual, *AIC* Akaike information criterion

Table 4 Parameters of the response shift model (M6)

Functioning scale	Regression weights (reprioritization) pre/post/then	Intercepts (uniform recalibration) pre/post/then	Residual variances (non-uniform recalibration) pre/post/then
PF	9.6/9.6/9.6	74.6/74.6/**67.3**	221.2/221.2/**292.4**
RF	17.7/17.7/17.7	55.9/55.9/55.9	545.1/545.1/545.1
SF	18.1/18.1/18.1	59.0/59.0/59.0	549.5/**292.1/292.1**
CF	20.7/**16.0/16.0**	70.8/**68.2**/70.8	327.9/327.9/327.9
EF	24.1/**17.9/17.9**	55.1/55.1/55.1	299.9/299.9/299.9

Bold are parameters that differed from pretest

Latent variables mean (standard deviation): pretest 0.00 (1.00), posttest 0.47 (1.28), thentest 0.11 (1.24) and latent variables correlations: r(pre, post) = 0.69, r(then, post) = 0.81, r(pre, then) = 0.75

PF Physical functioning, *RF* role functioning, *SF* social functioning, *CF* cognitive functioning, *EF* emotional functioning

latent variable that belongs to the shifted item (pretest: 0.11, posttest: 0.47). The effect does not occur in the perceived change comparison because the weights do not differ here. Regarding emotional functioning, the reprioritization effect decreased the observed change by 2.93 points and the systematic deviation between the thentest and the pretest by − 0.68 points. When comparing thentest and pretest responses (differences of recollection) and taking response shift into consideration (true change), a systematic deviation (recall bias) in one direction was revealed, that is, the patients seemed to remember their former functioning as having been slightly better than it actually was (raw differences below 2.63 points, effect sizes below 0.1).

Discussion

The first aim of this study was to compare HRQoL of cardiac patients with HRQoL of the general population. At baseline (pretest, during cardiac rehabilitation), the patients' HRQoL differed significantly on all scales. (All effect sizes were above 0.3.) Similar results have been reported by other studies, e.g., Juenger et al. [4] for all of the Short Form Health Survey SF-36 scales, and Schweikert et al. [5] for the *usual activities* scale amongst others of the EQ-5D health states. The two functioning scales that differed the most were role and social functioning. The largest differences in symptoms

Table 5 Decomposition of response shift effects (raw differences of implied means, range 0–100)

	Observed difference	True change	Response shift		Pooled SD for Hedges' *g*
			Uniform recalibration	Reprioritization	
Observed change (posttest–minus–pretest)					
PF	4.49	4.49	0.00	0.00	18.60
RF	8.26	8.26	0.00	0.00	31.06
SF	8.47	8.47	0.00	0.00	29.29
CF	4.87	9.66	−2.60	−2.20	27.46
EF	8.33	11.26	0.00	−2.93	29.25
Perceived change (posttest–minus–thentest)					
PF	10.71	3.44	7.27	0.00	20.18
RF	6.34	6.34	0.00	0.00	32.36
SF	6.49	6.49	0.00	0.00	28.60
CF	3.13	5.72	−2.60	0.00	27.15
EF	6.39	6.39	0.00	0.00	28.44
Difference of recollection (thentest–minus–pretest)					
PF	−6.22	1.05	−7.27	0.00	19.56
RF	1.93	1.93	0.00	0.00	30.83
SF	1.98	1.98	0.00	0.00	29.05
CF	1.74	2.25	0.00	−0.51	27.26
EF	1.94	2.63	0.00	−0.68	29.04

SD Standard deviation, *PF* physical functioning, *RF* role functioning, *SF* social functioning, *CF* cognitive functioning, *EF* emotional functioning

Example for interpretation: The observed difference is decomposed into the sum of true change (effect if no response shift has occurred), uniform recalibration and reprioritization, e.g., in CF (observed change) we observed a difference of 4.87 points that increased to 9.66 points when response shift was considered. On account of uniform recalibration, the effect was reduced by −2.60 points and because of reprioritization it was reduced by another −2.20 points

were found in the scales for fatigue and dyspnoea. Although patients' HRQoL had increased significantly three months after rehabilitation, their mean scores were still lower than those of the general population.

The second aim was to investigate how changes in HRQoL were observed and perceived. We took a closer look at the functioning scales and found that, over a 3-month period, the patients perceived changes in their HRQoL differently than they were observed. In the physical functioning domain, they perceived more change, and in the role, social, cognitive, and emotional functioning domains, they perceived less change. After taking response shift into consideration, the perceived changes on all functioning scales were lower than observed, whether using the thentest or the SEM approach.

The third aim of this study was to explore response shift effects and indications of recall bias more closely. We identified different kinds of response shift. *Uniform recalibration* affected physical functioning (thentest) and cognitive functioning (posttest). Regarding physical functioning, the patients judged their actual level of functioning (at pretest and posttest) in the same way, but when they retroactively assessed their former physical functioning (thentest) they reported lower scores. This result is in line with another study on patients undergoing cardiac rehabilitation [7]. It is possible that the social experience of cardiac rehabilitation plays a role whereby the patients come into contact to other patients with similar levels of physical functioning. They learned that their level of physical functioning was actually worse than they thought, started to cope with that, and finally recalibrated their internal values. At follow-up, they reported their recalibrated former physical functioning, but the value of their actual physical functioning did not change. Regarding cognitive functioning, the patients judged their former cognitive functioning equally at the pretest and the thentest, but they judged their current cognitive functioning (posttest) to be lower than before, e.g., 3 months after rehabilitation they felt to have more difficulties concentrating and remembering. *Reprioritization* affected cognitive functioning (pretest) and emotional functioning (pretest), indicating that at follow-up (posttest and thentest) the patients attached less importance to cognitive functioning and emotional functioning when they assessed their overall functioning. This means that cognitive and emotional functioning had less impact on their overall assessments. This might be due to diminishing cognitive and emotional strain as a result of

new knowledge acquired during cardiac rehabilitation. The patients may have learned new ways of responding to the emotional challenges of their illness and gained cognitive ability. Consequently, the strains on their cognitive and emotional functioning diminished and lost importance compared to other types of functioning. *Non-uniform recalibration* affected physical functioning (thentest) and social functioning (pretest). Lower residual variance (social functioning in post- and thentest) indicates less distance to the mean and suggests that the respondents answered in a more differentiated or more precise way in comparison with the pretest measurements. Higher residual variance (physical functioning in thentest) indicates an increase in random error due to less differentiated or less precise answers.

The indication of *recall bias* was marginal and did not influence the conclusion that was indicated by the thentest approach. The SEM results differed in some effects, but not due to recall bias. A convergent validity study [16] that also took both approaches into account did not find indications of relevant recall bias either. In the thentest approach, response shift is measured with the mean of the difference thentest-minus-pretest. It showed the only statistically significant effect in physical functioning ($g = -0.32$) that was also identified with the SEM approach (raw difference −7.27 points, $g = -0.37$). But the SEM approach revealed an additional recalibration effect, which is in line with another study [38]. The uniform recalibration effect in cognitive functioning of −2.60 points ($g = -0.10$) could not have been detected by the thentest approach, even without the occurrence of any recall bias. This is because this effect turned out only in the posttest measurement of cognitive functioning, which is not reflected in the thentest–minus–pretest difference. Consequently it seems obvious that the two approaches are not equivalent and do provide converging results only under very special circumstances (no reconceptualization, no reprioritization, no recalibration in the posttest measurement, minimal recall bias).

Limitations

We analyzed a more or less homogenous group of patients who underwent cardiac rehabilitation. On the one hand, the generalizability of the results to other kinds of diseases is unclear, but on the other hand all patients had a common catalyst that may explain these findings. A comparison with a control group of patients who are not undergoing cardiac rehabilitation could attribute the effects to the intervention. Furthermore, the EORTC QLQ-C30 is an instrument developed to assess HRQoL in patients with cancer. Thus, the recorded symptoms are those commonly reported by cancer patients [25]. On the other hand, we based the essential part of our analysis on the functioning scales, which contain the main components that define HRQoL, they are the functional effects of physical, mental, and social response to disease and treatment [39]. While disease-specific instruments have limited sensitivity in identifying differences between different groups, e.g., the comparison with the general population, they are still more sensitive to change than generic questionnaires [40].

Conclusions

In summary we found that cardiac patients have markedly worse HRQoL in all dimensions of the EORTC QLQ-C30, even 3 months after cardiac rehabilitation. We found that response shift effects *do* occur, something that should be taken into account when changes in HRQoL over time are studied. Simple post–pre differences can underestimate real changes. Furthermore, in the case of uniform recalibration (unequal intercepts) a comparison of the latent means of pre-, post- or thentest is not defensible because of the shifted metric of the latent variable. Combining both methods (thentest and structural equation modeling) proved to be essential for detecting more comprehensive evidence of response shift.

Acknowledgements We thank all patients who participated in this study.

Funding This study was supported by Deutsche Forschungsgemeinschaft (Grant Number: HI 1108/5-1).

Compliance with ethical standards

Conflicts interest The authors declare that they have no conflicts of interest.

Ethical approval The study received research ethics committee approval from ethic board of the medical faculty of the University of Leipzig (chairperson: Prof. Dr. R. Preiß, protocol number: 287-13-07102013, date of approval: 15 Oct 2013).

Informed consent Informed consent was obtained from all participants individual included in the study.

Research involving human participants and/or animals All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. The study received research ethics committee approval from ethic board of the medical faculty of the University of Leipzig (chairperson: Prof. Dr. R. Preiß, protocol number: 287-13-07102013, date of approval: 15 Oct 2013).

References

1. Rumsfeld, J. S., Alexander, K. P., Goff, D. C., Graham, M. M., Ho, P. M., Masoudi, F. A., et al. (2013). Cardiovascular health: The importance of measuring patient-reported health status. A

scientific statement from the American Heart Association. *Circulation, 127*(22), 2233–2249. https://doi.org/10.1161/cir.0b013e3182949a2e.
2. Landman, G. W. D., van Hateren, K. J. J., Kleefstra, N., Groenier, K. H., Gans, R. O. B., & Bilo, H. J. G. (2010). Health-related quality of life and mortality in a general and elderly population of patients with type 2 diabetes (ZODIAC-18). *Diabetes Care, 33*(11), 2378–2382. https://doi.org/10.2337/dc10-0979.
3. Lane, D. A., Lip, G. Y. H., & Millane, T. A. (2002). Quality of life in adults with congenital heart disease. *Congenital Heart Disease, 88*(1), 71. https://doi.org/10.1136/heart.88.1.71.
4. Juenger, J., Schellberg, D., Kraemer, S., Haustetter, A., Zugck, C., Herzog, W., et al. (2002). Health related quality of life in patients with congestive heart failure: Comparison with other chronic diseases and relation to functional variables. *Cardiovascular Medicine, 87*(3), 235–241. https://doi.org/10.1136/heart.87.3.235.
5. Schweikert, B., Hunger, M., Meisinger, C., König, H.-H., Gapp, O., & Holle, R. (2009). Quality of life several years after myocardial infarction: Comparing the MONICA/KORA registry to the general population. *European Heart Journal, 30*(4), 436–443. https://doi.org/10.1093/eurheartj/ehn509.
6. de Smedt, D., Clays, E., Annemans, L., Pardaens, S., Kotseva, K., & de Bacquer, D. (2015). Self-reported health status in coronary heart disease patients: A comparison with the general population. *European Journal of Cardiovascular Nursing, 14*(2), 117–125. https://doi.org/10.1177/1474515113519930.
7. Dempster, M., Carney, R., & McClements, R. (2010). Response shift in the assessment of quality of life among people attending cardiac rehabilitation. *British Journal of Health Psychology, 15*(Pt 2), 307–319. https://doi.org/10.1348/135910709X464443.
8. Deutsche Rentenversicherung Bund. (2016). Reha-Therapiestandards Koronare Herzkrankheit für die medizinische Rehabilitation der Rentenversicherung.
9. Howard, G. S., Ralph, K. M., Gulanick, N. A., Maxwell, S. E., Nance, D. W., & Gerber, S. K. (1979). Internal invalidity in pretest-posttest self-report evaluations and a re-evaluation of retrospective pretests. *Applied Psychological Measurement, 3*(1), 1–23. https://doi.org/10.1177/014662167900300101.
10. Howard, G. S., & Dailey, P. R. (1979). Response-shift bias: A source of contamination of self-report measures. *Journal of Applied Psychology, 64*(2), 144–150. https://doi.org/10.1037/0021-9010.64.2.144.
11. Sprangers, M. A. G., van Dam, F. S. A. M., Broersen, J., Lodder, L., Wever, L., Visser, M. R. M., et al. (1999). Revealing response shift in longitudinal research on fatigue: The use of the thentest approach. *Acta Oncologica, 38*(6), 709–718. https://doi.org/10.1080/028418699432860.
12. Barclay-Goddard, R., Epstein, J. D., & Mayo, N. E. (2009). Response shift: A brief overview and proposed research priorities. *Quality of Life Research, 18*(3), 335–346. https://doi.org/10.1007/s11136-009-9450-x.
13. Oort, F. J. (2005). Using structural equation modeling to detect response shifts and true change. *Quality of Life Research, 14*(3), 587–598. https://doi.org/10.1007/s11136-004-0830-y.
14. Sprangers, M. A. G., & Schwartz, C. E. (1999). Integrating response shift into health-related quality of life research: A theoretical model. *Social Science and Medicine, 48*(11), 1507–1515. https://doi.org/10.1016/S0277-9536(99)00045-3.
15. Schwartz, C. E., & Sprangers, M. A. G. (2010). Guidelines for improving the stringency of response shift research using the thentest. *Quality of Life Research, 19*(4), 455–464. https://doi.org/10.1007/s11136-010-9585-9.
16. Visser, M. R. M., Oort, F. J., & Sprangers, M. A. G. (2005). Methods to detect response shift in quality of life data: A convergent validity study. *Quality of Life Research, 14*(3), 629–639. https://doi.org/10.1007/s11136-004-2577-x.
17. Schwarz, R., & Hinz, A. (2001). Reference data for the quality of life questionnaire EORTC QLQ-C30 in the general German population. *European Journal of Cancer, 37*(11), 1345–1351. https://doi.org/10.1016/S0959-8049(00)00447-0.
18. Hinz, A., Singer, S., & Brähler, E. (2014). European reference values for the quality of life questionnaire EORTC QLQ-C30: Results of a German investigation and a summarizing analysis of six European general population normative studies. *Acta Oncologica, 53*(7), 958–965. https://doi.org/10.3109/0284186X.2013.879998.
19. Michelson, H., & Bolund, C. (2009). Health-related quality of life measured by the EORTC QLQ-C30: Reference values from a large sample of the Swedish population. *Acta Oncologica, 39*(4), 477–484. https://doi.org/10.1080/028418600750013384.
20. Derogar, M., van der Schaaf, M., & Lagergren, P. (2012). Reference values for the EORTC QLQ-C30 quality of life questionnaire in a random sample of the Swedish population. *Acta Oncologica, 51*(1), 10–16. https://doi.org/10.3109/0284186X.2011.614636.
21. Djärv, T., Wikman, A., & Lagergren, P. (2012). Number and burden of cardiovascular diseases in relation to health-related quality of life in a cross-sectional population-based cohort study. *British Medical Journal Open*. https://doi.org/10.1136/bmjopen-2012-001554.
22. Waldmann, A., Schubert, D., & Katalinic, A. (2013). Normative data of the EORTC QLQ-C30 for the German population: A population-based survey. *PLoS ONE, 8*(9), e74149. https://doi.org/10.1371/journal.pone.0074149.
23. Fredheim, O. M. S., Borchgrevink, P. C., Saltnes, T., & Kaasa, S. (2007). Validation and comparison of the health-related quality-of-life instruments EORTC QLQ-C30 and SF-36 in assessment of patients with chronic nonmalignant pain. *Journal of Pain and Symptom Management, 34*(6), 657–665. https://doi.org/10.1016/j.jpainsymman.2007.01.011.
24. Hasheesh, M. O. A., Almostafa, O. Y., & Ahmed, M. (2010). Health related quality of life among cardiac disease patients at Queen Alia Heart Institute. *Jordan Medical Journal,* 44.
25. Aaronson, N. K., Ahmedzai, S., Bergman, B., Bullinger, M., Cull, A., Duez, N. J., et al. (1993). The European Organization for Research and Treatment of Cancer QLQ-C30: A quality-of-life instrument for use in international clinical trials in oncology. *Journal of the National Cancer Institute, 1993*(85), 365–376.
26. Fayers, P. M., Aaronson, N. K., Bjordal, K., Groenvold, M., Curran, D., & Bottomley, A on behalf of the EORTC Quality of Life Group. (2001). *The EORTC QLQ-C30 Scoring Manual (3rd Edition)*. Brussels.
27. Giesinger, J. M., Kieffer, J. M., Fayers, P. M., Groenvold, M., Petersen, M. A., Scott, N. W., et al. (2016). Replication and validation of higher order models demonstrated that a summary score for the EORTC QLQ-C30 is robust. *Journal of Clinical Epidemiology, 69,* 79–88. https://doi.org/10.1016/j.jclinepi.2015.08.007.
28. Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B, (Statistical methodology),* 39(1), 1–38.
29. Zaiontz, C. (2015). *Real Statistics Using Excel*.
30. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: L. Erlbaum Associates; Taylor and Francis.
31. Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothenstein, H. R. (2009). *Introduction to meta-analysis*. New York, NY: John Wiley & Sons Ltd.
32. Oort, F. J., Visser, M. R. M., & Sprangers, M. A. G. (2005). An application of structural equation modeling to detect response shifts and true change in quality of life data from cancer patients

undergoing invasive surgery. *Quality of Life Research, 14*(3), 599–609. https://doi.org/10.1007/s11136-004-0831-x.
33. Gerlich, C., Schuler, M., Jelitte, M., Neuderth, S., Flentje, M., Graefen, M., et al. (2016). Prostate cancer patients' quality of life assessments across the primary treatment trajectory: 'True' change or response shift? *Acta Oncologica, 55*(7), 814–820. https://doi.org/10.3109/0284186X.2015.1136749.
34. Saris, W. E., Satorra, A., & van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling, 16*(4), 561–582. https://doi.org/10.1080/10705510903203433.
35. Cocks, K., King, M. T., Velikova, G., de Castro, G., Martyn St-James, M., Fayers, P. M., et al. (2012). Evidence-based guidelines for interpreting change scores for the European Organisation for the Research and Treatment of Cancer Quality of Life Questionnaire Core 30. *European Journal of Cancer Care, 48*(11), 1713–1721. https://doi.org/10.1016/j.ejca.2012.02.059.
36. Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118.
37. Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology, 93*(3), 568–592. https://doi.org/10.1037/0021-9010.93.3.568.
38. Piwowar, V., & Thiel, F. (2014). Evaluating response shift in training evaluation: Comparing the retrospective pretest with an adapted measurement invariance approach in a classroom management training program. *Evaluation Review, 38*(5), 420–448. https://doi.org/10.1177/0193841X14546932.
39. Gierlaszyńska, K., Pudlo, R., Jaworska, I., Byrczek-Godula, K., & Gąsior, M. (2016). Tools for assessing quality of life in cardiology and cardiac surgery. *Kardiochirurgia i torakochirurgia polska = Polish journal of cardio-thoracic surgery, 13*(1), 78–82. https://doi.org/10.5114/kitp.2016.58974.
40. Briançon, S., Gergonne, B., Guillemin, F., Empereur, F., & Klein, S. (2002). Disease-specific versus generic measurement of health-related quality of life in cross-sectional and longitudinal studies: An inpatient investigation of the SF-36 and four disease-specific instruments. In M. Mesbah, B. F. Cole, & M.-L. T. Lee (Eds.), *Statistical methods for quality of life studies: Design, measurements and analysis* (pp. 87–99). Boston, MA: Springer.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.