Using structural equation modeling to investigate change in health-related quality of life

Verdam, M.G.E.

Publication date
2017

Document Version
Other version

License
Other

CHAPTER 2

Response Shift Detection Through Then-Test and Structural Equation Modeling: Decomposing Observed Change and Testing Tacit Assumptions

Assessment of change in patient-reported outcomes may be invalidated by the occurrence of response shift. Response shift refers to a change in respondent's frame of reference that may cause changes in observed variables that are not directly related to change in the construct of interest. An established approach for detecting response shift in the area of health-related quality of life (HRQL) is to administer a retrospective pre-test (then-test). In this study, the then-test was incorporated in the structural equation modeling (SEM) approach to (1) compare the then-test approach and the SEM approach in their decomposition of observed change and (2) to test the underlying assumptions of the then-test approach. In an application to HRQL-data of 170 cancer patients undergoing invasive surgery, we found that both approaches revealed a similar pattern of decomposition, although there were some differences in the size and direction of change. With regard to the underlying assumptions of the then-test approach, results showed: (1) no evidence for recall-bias (Recall Assumption supported for all scales), (2) that frames of reference were not invariant across post- and then-test measures (Consistency Assumption rejected for four out of nine scales), and (3) that frames of reference were not only affected by the recalibration type of response shift (Recalibration Assumption rejected for three out of nine scales). Future research should focus on valid approaches for detecting response shift and the consequences for assessing changes in HRQL.

This chapter is based on: Verdam, M. G. E., Oort, F. J., Visser, M. R. M., & Sprangers, M. A. G. (2012). Response shift detection through then-test and structural equation modelling: Decomposing observed change and testing tacit assumptions. *Netherlands Journal of Psychology, 67,* 58-67.

Introduction

Patient-reported outcomes of health-related quality of life (HRQL) are becoming increasingly more important in evaluating treatment effects in clinical settings. However, there is a well-known disparity between patient-reported and clinical measures of function. One explanation for this disparity is related to the dynamic nature of the HRQL construct (Allison, Locker, & Feine, 1997). The dynamic nature of the construct entails that the frame of reference with which individuals assess their HRQL can differ between subjects and can change within subjects over time. Such a change in frame of reference may cause changes in observed variables that are not directly related to change in the construct of interest. It is therefore important to detect possible changes in respondent's frame of reference.

Change in frames of reference is also referred to as "response shift". The term response shift was first used in research on educational training interventions (Howard et al., 1979) and was also investigated in the field of organizational change where they used the terminology of "alpha", "beta" and "gamma" change (Golembiewski, Billingsley, & Yeager, 1976). In the area of HRQL-research, Schwartz & Sprangers (1999) proposed a theoretical model of response shift that distinguishes three types of response shift: (1) recalibration, which refers to a change in the respondent's internal standards of measurement, (2) reprioritization, that refers to a change in respondent's values regarding the relative importance of component domains of the target construct, and (3) reconceptualization, referring to a change in definition of the target construct. Response shift causes comparison of measurements over time to be incomparable. Therefore, when investigating changes in HRQL, it is important to also investigate – and account for – response shift effects.

Several methodological approaches are available to investigate response shift in longitudinal HRQL-research (Schwartz & Sprangers, 1999; Schwartz et al., 2011). The 'then-test' approach is most commonly used, and includes a retrospective pre-test measure in addition to the usual pre- and post-measures. This retrospective pre-test is administered at the post-test occasion and asks respondents to re-evaluate their HRQL at the time of pre-test. As the then-test and post-test are administered at the same time, it is assumed that both measurements are completed with the same frame of reference, thus avoiding response shift effects. Comparison of the post-test and then-test scores would yield an unbiased indication of the treatment effect ('true' change, see Table 1). Furthermore, differences between the then-test and pre-test scores could be used as an assessment of changes in subjects' frames of reference (response shift). The then-test approach thus allows a decomposition of observed change (differences between pre-test and post-test scores) into true change and response shift. However, these interpretations are only valid when the following assumptions are met:

(1) Recall Assumption: At then-test occasion respondents are able to recall their state at pre-test. The validity of the then-test depends on the underlying assumption that memory (the recall of the pre-test state) is accurate and alternative cognitive explanations (e.g.

social desirability, cognitive dissonance, implicit theory of change, expectancy or experimenter effects) do not play a role.

(2) Consistency Assumption: Post- and then-test are completed with the same frame of reference. A valid comparison of then-test and post-test scores depends on the underlying assumption that the respondents' frames of reference are invariant across these assessments.

(3) Recalibration Assumption: All response shift is of the recalibration type. As the then-test approach aims to assess only recalibration – not reprioritization and reconceptualization – the comparison of then-test and pre-test scores in assessing response shift is only accurate if all response shift is of the recalibration type.

An alternative method to detecting response shift is the structural equation modeling (SEM) approach (Oort, 2005). Similar to the then-test approach, the SEM-approach provides a way to decompose observed change into true change and response shift (Oort, 2005, p. 495), based on the estimates of the factor model parameters (see Table 1). An advantage of the SEM approach is that it allows for the statistical comparison of separate components of the measurement model over time, enabling operationalization of the different types of response shift.

Table 1 | Decomposition of observed change according to the then-test approach and the SEM approach

Then-test approach	
Observed change = True change	+ Recalibration
$(X_{post} - X_{pre}) = (X_{post} - X_{then})$	$+ (X_{then} - X_{pre})$
SEM approach	
Observed change = True change	+ Recalibration + Reprioritization & Reconceptualization
$(\mu_{post} - \mu_{pre}) = \Lambda_{pre}{}^{*} (\kappa_{post} - \kappa_{pre}) + (\tau_{post} - \tau_{pre})$	$+ (\Lambda_{post} - \Lambda_{pre}) * \kappa_{post}$

Notes: In the then-test approach scores for the different measurements are denoted with 'X' to reflect the observed nature of the scores. In the SEM-approach Greek symbols reflect the parameter estimates of observed factor means (μ), common factor loadings (Λ), common factor means (κ) and intercepts (τ).

The SEM approach can therefore be used not only as a technique for the detection of response shift, but also for a substantive analysis of the decomposition of change. Moreover, the characteristics of the SEM approach provide a unique opportunity to test the underlying assumptions of the then-test approach. Incorporating the then-test into the SEM approach allows for testing the validity (and consistency) of the measurement model for post- and then-test (Consistency Assumption) and assessing not only the occurrence of recalibration, but also reprioritization and reconceptualization (Recalibration Assumption). Moreover, recall bias can be investigated by examining effects on the underlying constructs instead of the observed variables (Recall Assumption).

Therefore, the aim of this study is to illustrate how incorporation of the then-test into the SEM approach enables: (1) a substantive comparison of both approaches in their decomposition of observed change into true change and (types of) response shift, and (2) testing the underlying assumptions of the then-test approach.

Method

Cancer patients' health-related quality of life was assessed prior to surgery (pre-test) and three months following surgery (post-test and then-test). These data have been used before to investigate response shift with the then-test and the SEM approach (Visser, Oort, & Sprangers, 2005).

Patients

A consecutive series of 170 newly diagnosed cancer patients were enrolled, including 29 lung cancer patients undergoing either lobectomy or pneumectomy, 43 pancreatic cancer patients undergoing pylorus-preserving pancreaticoduodenectomy, 46 esophageal cancer patients undergoing either transhiatal or transthoracic resection and 52 cervical cancer patients undergoing hysterectomy. Exclusion criteria were being under the age of 18, having a life expectancy less than 9 months, or not being able to complete a (Dutch) questionnaire. The sample consisted of 87 men and 83 women, with ages ranging from 27 to 83 (mean 57.5, standard deviation 14.1).

Measures

Generic health-related quality of life was assessed with the Dutch language version (Aaronson et al., 1998) of the SF-36 health survey (Ware, Snow, Kosinski, & Gandek, 1993), encompassing eight scales: physical functioning (PF), role limitations due to physical health (role-physical, RP), bodily pain (BP), general health perceptions (GH), vitality (VT), social functioning (SF), role limitations due to emotional problems (role-emotional, RE), and mental health (MH). Fatigue (FT) was measured with a six-item short form of the multidimensional fatigue inventory (MFI; Smets, Garssen, Bonke, & De Haes, 1995), to cover effects on patients' fatigue more thoroughly. For computational convenience the original scale scores of the SF-36 scales and the short form of the MFI were transformed to scales ranging from 0 to 5, with higher scores indicating better health. There were no missing data, as completion of the self-administered questionnaires was checked by an interviewer.

Structural Equation Modeling

The SEM procedure (Oort, 2005) was applied to the data of pre-, post- and then-tests to detect response shift and includes: (1) establishing an appropriate measurement model, (2) fitting a model of no response shift, (3) detection of response shift, and (4) assessment of true change. The measurement model was established on the basis of published results of principal components analyses of the SF-36 (Ware et al., 1993), results of exploratory factor analyses of the present data, and substantive considerations. The measurement model has no across measurement constraints. To test for the occurrence of response shift the second step in the SEM procedure is to fit a model of no response shift (where all model parameters that are associated with response shift are constrained to be equal across measurements). To test for the presence of response shift,

the no response shift model is compared to the model with no across measurement constraints. The third step in the SEM procedure begins with the no response shift model and uses step-by-step modification to arrive at the response shift model where all apparent response shifts are accounted for. Response shift is operationalized as across-measurement differences between patterns of common factor loadings (reconceptualization), values of common factor loadings (reprioritization), differences between intercepts (uniform recalibration), and differences between residual variances (nonuniform recalibration). In the fourth step of the SEM procedure, true change is assessed in the model where response shift is accounted for.

Structural equation models were fitted to the means, variances and covariances of the SF-36 and MFI scale scores of pre-, post- an then-test, using standard statistical computer programs (Jöreskog & Sorbom, 1996; Neale, Boker, Xie, & Maes, 1999) (LISREL provides modification indices and Mx provides likelihood-based confidence intervals). To achieve identification of all model parameters, scales and origins of the common factors were established by fixing the factor means at zero and the factor variances at one. In Steps 2 and 3 of the procedure, only first occasion (pre-test) factor means and variances are fixed; post-test and then-test factor means and variances are then identified by constraining intercepts and common factor loadings to be equal across assessments (Oort, 2005).

Goodness-of-fit was evaluated with the chis-square test of exact fit (CHISQ), where a significant chi-square indicates a significant difference between data and model. However, in the practice of structural equation modeling, exact fit is rare, and with large sample sizes the chi-square test generally turns out to be significant. An alternative measure of overall goodness-of-fit is the root mean square error of approximation (RMSEA). According to a generally accepted rule of thumb, an RMSEA value below .08 indicates 'reasonable' fit and one below .05 'close' fit (Browne & Cudeck, 1992). In addition, the comparative fit index (CFI; Bentler, 1990) gives an indication of model fit based on model comparison (compared to the independence model in which all measured variables are uncorrelated), where CFI of .97 or higher is indicative of good fit and CFI between .95 and .97 of acceptable fit. Yet another fit index is the expected cross validation index (ECVI; Browne & Cudeck, 1989) which is a measure of the discrepancy between the model-implied covariance matrix in the analyzed sample ('calibration' sample), and the covariance matrix that would be expected in another sample of the same size ('validation' sample). The ECVI can be used to compare different models for the same data, where the model with the smallest ECVI indicates the model with the best fit.

The chi-square difference test ($CHISQ_{diff}$; Bollen, 1989) was used to compare the fit of nested models, where a significant chi-square indicates that the addition of model parameters significantly improves the model fit. Significant modification indices (Jöreskog & Sorbom, 1996) and standardized residuals > .10 were assumed to indicate response shift. The specification search was consistently guided by substantive consideration in order to retain a theoretical sensible model. Each modification was tested with the $CHISQ_{diff}$.

Objective 1: Decomposition of Change

Equations in Table 1 give the decomposition of observed change into true change and response shifts for both the then-test approach and the SEM approach. For the then-test approach the standard deviations of the observed change scores are used to calculate standardized mean differences (as effect size indices *d*) for the components of observed change. For the SEM approach the parameter estimates of the final model (in which all response shifts are accounted for) were used to calculate standardized mean differences (as effect size indices *d*) for the components of observed change (Oort, 2005). Effect-size values of $d = .2, .5$ and .8 are considered 'small', 'medium', and 'large' (Cohen, 1988).

Objective 2: Testing the Assumptions of the Then-Test Approach

The Recall Assumption can be tested by testing the equality of pre-test and then-test common factor means because the common factor means of the response shift model should refer to the same state (of pre-test). The Recall Assumption would be supported when the equality constraint across pre- and then-test common factor means is tenable (indicated by the $CHISQ_{diff}$).

The Consistency Assumption can be tested by imposing equality constraints across post- and then-test common factor loadings (reconceptualization and reprioritization), intercepts (uniform recalibration) and residual variances (nonuniform recalibration). When response shift detection (using the $CHISQ_{diff}$) is invariant across assessments, the Consistency Assumption is supported.

The Recalibration Assumption can be tested by examining recalibration, reprioritization and reconceptualization types of response shift. When all response shifts detected (using the $CHISQ_{diff}$) are of the recalibration type, the Recalibration Assumption is supported.

Results

Table 2 gives pre-, post- and then-test means and standard deviations for all SF-36 and MFI scales.

Table 2 | Means and standard deviations for SF-36 and MFI scales before surgery (pre-test) and three months after surgery (post-test and then-test)

	Pre-test		Post-test		Then-test	
Scale	Mean	SD	Mean	SD	Mean	SD
PF	3.96	1.22	3.18	1.32	4.05	1.37
RP	2.73	2.09	2.13	2.02	2.99	2.14
BP	3.94	1.19	3.68	1.21	4.20	1.27
SF	3.81	1.32	3.62	1.47	3.72	1.32
MH	3.25	1.08	3.69	1.05	3.26	1.14
RE	3.00	2.12	3.55	1.93	2.84	2.13
VT	3.14	1.26	2.77	1.23	3.18	1.32
GH	2.96	0.95	2.96	1.06	2.76	1.08
FT	3.30	1.10	2.92	1.18	3.24	1.17

Notes: $N = 170$; SF-36 and MFI scale scores range from 0 to 5.

Measurement Model

Results from exploratory factor analyses and substantive considerations gave rise to the measurement model in Figure 1 (see Oort, Visser, & Sprangers, 2005 for more information on selection of this measurement model). The circles represent unobserved, latent variables and the squares represent the observed variables. Three latent variables are the common factors general physical health (GenPhys), general mental health (GenMent), and general fitness (GenFitn). GenPhys is measured by PF, RP, BP and SF, GenMent is measured by MH, RE, and again SF, and GenFitn is measured by VT, GH, and FT. Other latent variables are the residual factors ResPF, ResRP, ResBP, etc. The residual factors represent all that is specific to PF, RP, BP, etc., plus random error variation. In addition, Figure 1 shows the response shift model, the model in which all response shifts are accounted for (dotted lines represent common factor loadings that were present at post- and/or then-test only). Numbers in Figure 1 are maximum likelihood estimates of common factor loadings, common factor correlations, residual variances, and three residual correlations (single values represent estimates that are constrained to be equal across pre-, post- and then-test, whereas multiple values represent separate estimates for pre-test (black), post-test (red), and then-test (blue)). Figure 2 gives a visual representation of the full longitudinal model that was fitted to the data.

The measurement model of Figure 1 was the basis for a structural equation model for pre-, post and then-test with no across measurement constraints. The chi-square test of exact fit was significant (CHISQ(255) = 349.13, $p<.001$) but the RMSEA measure indicated close fit (RMSEA = .041, see Table 3).

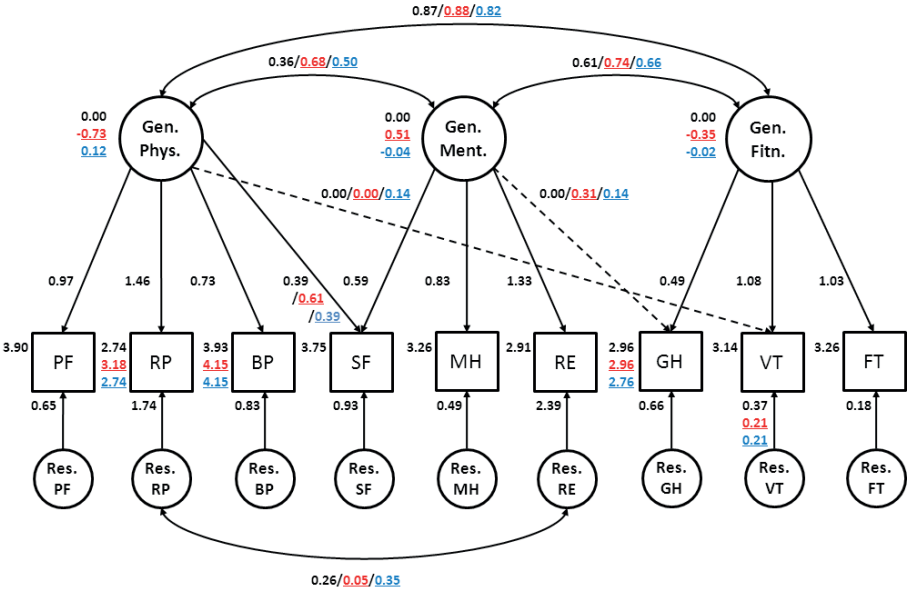

Figure 1 | The measurement model used in response shift detection

Notes: Circles represent latent variables (common and residual factors) and squares represent observed variables (the SF-36 and MFI scales). Numbers are maximum likelihood estimates of the response shift model parameters: common factor loadings, common factor correlations, residual variances, and a residual correlations. Single values represent estimates that were constrained to be equal across time, whereas multiple values represent different pre-test (black), post-test (red) and then-test (blue) estimates.

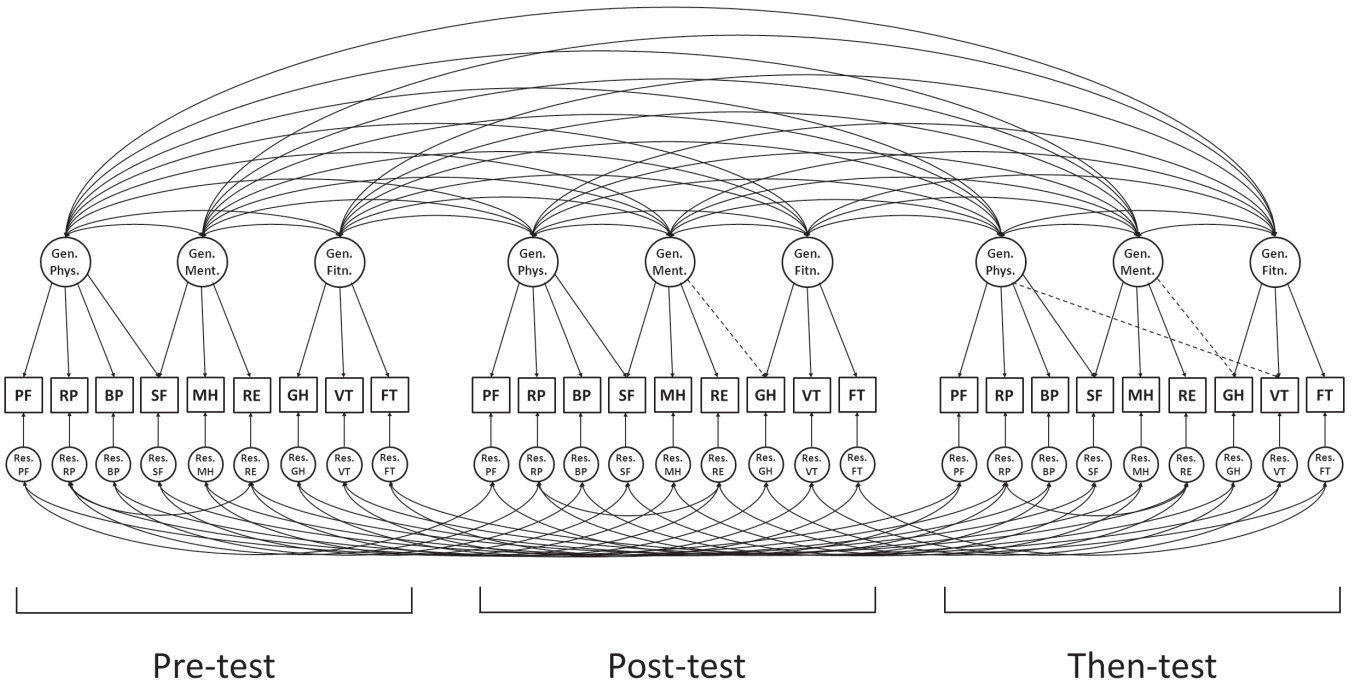

Figure 2 | The longitudinal structural equation model fitted to the data

Notes: Circles represent latent variables (common and residual factors) and squares represent observed variables (the SF-36 and MFI scales).

Dotted lines represent factor-loadings unique for post- or then-test assessment.

Detection of Response Shift

To test for the occurrence of response shift, all model parameters that are associated with response shift were held invariant across measurements. This means that all across measurement invariance constraints on common factor loadings, intercepts, and residual variances were imposed. The fit of the no response shift model, although still satisfactory (RMSEA = .049, see Table 3), was significantly worse than the fit of model with no across measurement constraints (chi-square difference test: $CHISQ_{diff}(44) = 99.26, p < .001$), indicating the presence of response shift.

Inspection of modification indices and standardized residuals indicated which of the equality constraints were not tenable. Step by step modifications yielded the response shift model, which showed several cases of response shift, as will be explained below. The fit of the response shift model was good (RMSEA = .035, see Table 3), and significantly better than the fit of the no response shift model ($CHISQ_{diff}(8) = 74.12, p < .001$). All estimates of the response shift model parameters are given in Table 4.

Table 3 | Goodness of overall fit of models in the three-step response shift detection procedure

Model	Description	DF	CHISQ	RMSEA	ECVI	CFI
Model 1	Measurement model (no across measurement constraints)	255	349.13	.041 (.026; .053)	3.39 (3.14; 3.69)	.99
Model 2	No response shift model	299	448.39	.049 (.037; .059)	3.73 (3.28; 3.92)	.98
Model 3	Response shift model	291	374.27	.035 (.019; .048)	3.43 (3.01; 3.58)	.99

Notes: $N = 170$; Numbers between parentheses represent 90% confidence intervals.

Evaluation of Response Shifts and True Change

Reconceptualization. A change in the pattern of common factor loadings across assessments is indicative of reconceptualization. Comparison of the common factor loadings of the pre-test with those of the post-test and then-test (Table 4, top rows) showed that at both the post- and then-test GH became an indicator for GenMent, indicating reconceptualization of GH. The VT scale became indicative of GenPhys at the then-test, indicating reconceptualization of VT at the then-test only.

Reprioritization. The values of the common factor loadings contain information about reprioritization. The common factor loading of SF on GenPhys became larger at the post-test, indicating reprioritization of SF at the post-test only.

Table 4 | Parameter estimates in the response shift model

	Pre-test			**Post-test**			**Then-test**		
	GenPhys	**GenMent**	**GenFitn**	**GenPhys**	**GenMent**	**GenFitn**	**GenPhys**	**GenMent**	**GenFitn**
Common factor loadings (Λ)									
PF	0.97			0.97			0.97		
RP	1.46			1.46			1.46		
BP	0.73			0.73			0.73		
SF	0.39	0.59		**0.61**	0.59		0.39	0.59	
MH		0.83			0.83			0.83	
RE		1.33			1.33			1.33	
VT			1.08			1.08	**0.14**		1.08
GH			0.49		**0.31**	0.49		**0.14**	0.49
FT			1.03			1.03			1.03
Intercepts (τ)									
	PF	RP	BP	SF	MH	RE	VT	GH	FT
Pre-test	3.90	2.74	3.93	3.75	3.26	2.91	3.14	2.96	3.26
Post-test	3.90	**3.18**	**4.15**	3.75	3.26	2.91	3.14	2.96	3.26
Then-test	3.90	2.74	**4.15**	3.75	3.26	2.91	3.14	**2.76**	3.26
Residual variance (Diag(Θ))									
	ResPF	ResRP	ResBP	ResSF	ResMH	ResRE	ResVT	ResGH	ResFT
Pre-test	0.65	1.74	0.83	0.93	0.49	2.39	0.37	0.66	0.18
Post-test	0.65	1.74	0.83	0.93	0.49	2.39	**0.21**	0.66	0.18
Then-test	0.65	1.74	0.83	0.93	0.49	2.39	**0.21**	0.66	0.18
Residual correlations (Θ)*									
Pre x Post	0.28	0.13	0.35	0.05	0.43	0.00	0.27	0.32	0.15
Pre x Then	0.62	0.22	0.42	0.26	0.54	-0.06	0.26	0.27	-0.02
Post x Then	0.41	0.04	0.19	0.18	0.58	0.26	0.26	0.27	0.22
Common factor variances (Diag(Φ))									
	Pre-test			Post-test			Then-test		
	GenPhys	GenMent	GenFitn	GenPhys	GenMent	GenFitn	GenPhys	GenMent	GenFitn
	1.00	1.00		1.23	0.86	1.13	1.33	1.19	1.08
Common factor correlations (Φ)*									
Pre-test									
Gen-Phys	1								
Gen-Ment	0.36	1							
Gen-Fitn	0.87	0.61							

Post-test								
Gen-Phys	0.55	0.35	1					
Gen-Ment	0.38	0.41	0.68	1				
Gen-Fitn	0.47	0.43	0.88	0.74	1			
Then-test								
Gen-Phys	0.82	0.37	0.41	0.25	0.32	1		
Gen-Ment	0.40	0.59	0.20	0.32	0.18	0.50	1	
Gen-Fitn	0.76	0.50	0.35	0.32	0.38	0.82	0.66	1
Common factor means (κ)								
	0.00	0.00	**-0.73**	**0.51**	**-0.35**	0.12	-0.04	-0.02

Notes: N = 170; Results indicating across-measurement variance are printed in bold. Common factor loadings are unstandardized, but covariances are decomposed into variances and correlations.

Recalibration. Intercepts and residual variances contain information about uniform and nonuniform recalibration. For RP, we found differences between the pre- and post-test intercepts, indicating uniform recalibration of RP at the post-test only. For GH, we found differences between the pre- and then-test intercepts, indicating uniform recalibration of GH at the then-test only. For BP, we found differences between the pre-test and both the post- and then-test intercepts, indicating uniform calibration of BP that equally affects both the post- and then-test. We also found a change in the variance of the residual factor ResVT, indicating nonuniform recalibration of VT that affects both the post- and then-test equally.

True change. Common factor means were fixed at zero for the pre-test (because of identification requirements), so that the post-test estimates serve as direct representations of true change. The differences between the pre- and post-test common factor means were significant ($p < .001$) for each of the common factors. GenPhys (-0.73) and GenFitn (-0.35) deteriorated, and GenMent (+0.51) improved, with effect-sizes that can be considered 'small' to 'medium' (d = -0.72, -0.37, and +0.48 respectively).

Objective 1: Comparison of Then-Test Approach and SEM Approach in the Decomposition of Observed Change

The results of the decomposition of observed change for both the then-test approach and the SEM approach are presented in Table 5.

Observed change. The results of the observed change indicate deteriorations that are considered 'small' effects for RP, BP, VT and FT, deterioration that is considered a 'medium' effect on PF and improvements that are considered 'small' effects for MH and RE. The pattern of observed change are found to be similar for both approaches, with only small differences in the standardized mean differences.

True change. Both the then-test approach and the SEM-approach also revealed a similar pattern of change for true change, except for GH. While the observed change for GH was not significant, both approaches revealed a significant true change for GH, albeit in the opposite direction. The then-test approach showed significant improvement of GH, while the SEM approach showed significant deterioration of GH.

Response shift. Both approaches revealed a significant positive response shift for BP, indicating that the true change of BP is larger than the observed change in BP. Only the SEM approach revealed a significant positive response shift for RP (resulting in a larger true change), and a significant negative response shift for SF (resulting in a smaller true change), while for the then-test approach these response shifts did not reach statistical significance. The response shifts detected for GH are in the opposite direction, with a negative response shift for GH according to the then-test approach and a positive response shift for GH according to the SEM approach, where the latter reaches a higher level of significance.

Table 5 | The decomposition of observed change into true change and response shift (displayed as standardized differences), for both the then-test approach and the SEM approach

	Then-test approach			SEM approach		
Scale	Observed change	True change	Response shift [a]	Observed change	True change	Response shift
PF	-0.59**	-0.66**	0.07	-0.51**	-0.51**	-
RP	-0.27**	-0.38**	0.12	-0.28**	-0.47**	0.19[a]**
BP	-0.20**	-0.40**	0.20**	-0.25**	-0.42**	0.17[a]**
SF	-0.11	-0.06	-0.05	-0.09	0.01	-0.10[b]*
MH	0.40**	0.39**	0.01	0.37**	0.37**	-
RE	0.21**	0.27**	-0.06	0.26**	0.26**	-
VT	-0.30**	-0.33**	0.03	-0.31**	-0.31**	-
GH	-0.00	0.18*	-0.18*	-0.01	-0.15**	0.14[c]**
FT	-0.35**	-0.30**	-0.05	-0.32**	-0.32**	-

Notes: N = 170; Standardized mean differences of 0.2, 0.5, and 0.8 indicate small, medium, and large differences (Cohen, 1988); *p < 0.05, **p < 0.01 in paired t-test (then-test approach) or inspection of confidence intervals (SEM approach); [a] = recalibration, [b] = reprioritization, [c] = reconceptualization.

Objective 2: Tenability of Assumptions Underlying the Then-Test Approach

Recall Assumption. Results indicate that the differences between pre-test and then-test common factor means were non-significant (p > .05) for all common factors: GenPhys (0.12), GenMent (-0.04), and GenFitn (-0.02).This indicates that the assumption that respondents are able to recall their state at pre-test has been met for all SF-36 and MFI scales.

Consistency Assumption. Results indicate that there are some parameters of the measurement model that are not invariant across post- and then-test measures: uniform recalibration of RP and reprioritization of SF on GenPhys affect only the post-test, while uniform calibration of GH and reconceptualization of VT for Genphys affect only the then-test. Also, the common factor loading of GH on GenMent differs between the post- and then-test. Therefore, the second assumption is rejected for GH, RP, SF and VT.

Recalibration Assumption. Results show that indeed some response shifts of the recalibration type were found (uniform recalibration of RP, BP and GH and nonuniform recalibration of VT), but that reprioritization (of SF) and reconceptualization (of GH on GenMent and VT on GenPhys) were also found. Therefore, the third assumption is rejected for GH, SF and VT.

Discussion

In this study a comparison was made between the then-test approach and the SEM approach in the detection of response shift in HRQL data from cancer patients undergoing invasive surgery. Results indicate that the decomposition of observed change is similar for both approaches, in that the size of true change is equal except for the direction of change in GH. The assessment of response shift differs somewhat, as only the SEM approach reveals response shifts for RP and SF, and the response shift detected in GH reaches a higher level of significance (see Visser et al., 2005 for a substantive explanation of these differences). In a study by Ahmed, Mayo, Wood-Dauphinee, Hanley, and Cohen (2005) the then-test approach was also compared to a method that also uses SEM. They did not detect any response shift using the SEM technique, while the then-test approach did reveal several response shifts. However, an explanation for this discrepancy could be that the measurement model used in the study by Ahmed et al. was suboptimal (Borsboom, Korfage, Essink-Bot, & Duivenvoorden, 2007) and that their SEM method is not as sensitive in detecting response shift effects as our SEM approach (Ahmed, Bourbeau, Maltais, & Mansour, 2009). In the present study, we showed that it is possible to use the SEM approach to make a substantive comparison between different methodologies for the detection of response shift by looking at the decomposition of observed change into true change and response shifts.

The second objective of this study was to test the underlying assumptions of the then-test approach. Our results supported the Recall Assumption for all scales (indicating no evidence of recall bias or alternative cognitive explanations), but failed to support the assumption that frames of reference are invariant across post-and then-test (Consistency Assumption rejected for four scales), and indicated that not all response shifts found were of the recalibration type (Recalibration Assumption rejected for three scales). These results are in line with a study of Nolte, Elsworth, Sinclair, and Osborne (2009) who applied SEM to assess psychometric properties of the then-test (using the health education impact questionnaire (heiQ)). They

tested measurement invariance for the pre- and post-test factor model and then- and post-test factor model. They found different types of response shift for the post- and the then-test, thus rejecting the Consistency Assumption for several scales, and concluded that the application of the then-test is not supported. Although their SEM approach used two models to test the underlying assumptions of the then-test approach, whereas our SEM approach consisted of a single combined model, both studies are illustrative of how the then-test can be incorporated into the SEM approach so that the underlying assumptions of the then-test approach can be evaluated. Testing the underlying assumptions of the then-test approach through SEM is useful for determining the validity of the then-test approach in assessing changes in HRQL.

If we combine our findings, we can assess the consequences of rejection of the assumptions underlying the then-test approach for the decomposition of observed change. For example, the rejection of the Recalibration Assumption (required for a valid assessment of response shift) for GH and SF coincides with a difference in assessment of response shift between the then-test approach and the SEM approach for these scales. Also, the rejection of the Consistency Assumption (required for a valid assessment of true change) for GH goes together with a difference in the assessment of true change between approaches, in that the then-test approach reveals a change in the opposite direction compared to the change detected in the SEM-approach. However, the rejections of the Consistency Assumption for RP and SF do not seem to affect the assessment of true change and rejection of assumptions for VT was inconsequential for the decomposition of observed change. Concluding, the rejection of underlying assumptions is reflected in the decomposition of observed change, but the pattern is not fully consistent.

The then-test approach and SEM approach use different methods for the detection of response shift. An advantage of the then-test approach is the relatively simple analysis for detecting response shift (e.g. t-tests). However, a valid assessment of change depends on the tenability of the Recall-, Recalibration-, and Consistency Assumptions. Also, the then-test approach requires an additional assessment, which can be an extra burden to patients. Using the SEM approach, there is no need for an additional assessment and a valid assessment of change does not depend on the tenability of the Recall-, Recalibration-, and Consistency Assumptions. However, the statistical analysis for the detection of response shift is relatively complicated. Moreover, decisions on which parameters are freed (e.g., which variable shows which type of response shift) are guided not only by statistical procedures or thresholds, but also by substantive considerations. This is necessary because relying on statistics alone could lead to freeing parameters that might not be theoretical sensible (e.g., an observed variable at post-test that is an indicator of a latent variable at pre-test). Consequently, this decision involves a subjective judgment of the researcher. For example, it could be that freeing the common factor loading of either one of two indicators of a latent variable yields the same result, and renders freeing the other common factor loading unnecessary. This means that the researchers has to decide which of those common factor loadings would be justified to free. The advantage is that response shift detection in the SEM approach is not only statistically but also theoretically driven, and will

therefore lead to more logical and probable models. A disadvantage is that results depend – partly – on these subjective decisions; others may make different choices. Therefore, it should be noted that it was not the objective of this study to draw substantive (clinical) conclusions about the response shifts found. The results in this study serve for illustrative purposes only.

In conclusion, incorporating the then-test into the SEM-approach: (1) allows for a comparison of the then-test approach and the SEM approach in their decomposition of observed change, (2) provides the possibility to test the underlying assumptions of the then-test approach, and (3) gives an idea of the consequences of rejection of underlying assumptions on the decomposition of observed change. To be able to draw valid conclusions in the assessment of HRQL, we need to be aware of the limitations of HRQL measurement. Quantifying the existence and size of response shifts and true change will help to better understand the observed change of HRQL. Future research should focus not only on validating the measurements of HRQL, but also on investigating the (clinical) consequences of violating the validity of change assessments.