UvA-DARE (Digital Academic Repository)

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

[Link to publication]

Citation for published version (APA):
Verdam, M. G. E. (2017). *Using structural equation modeling to investigate change in health-related quality of life*. [Thesis, fully internal, Universiteit van Amsterdam].

CHAPTER 8

The Impact of Response Shift on the Assessment of Change:
Calculation of Effect-Size Indices
Using Structural Equation Modeling

he investigation of response shift in patient-reported outcomes (PROs) is important in both clinical practice and research. Insight into the presence and strength of response shift effects is necessary for a valid interpretation of change in PROs. In this paper we illustrate how to evaluate the impact of response shift on the assessment of change through the calculation of effect-size indices of change. Specifically, when response shift is investigated through structural equation modeling (SEM), observed change can be decomposed into: 1) change due to recalibration response shift, 2) change due to reprioritization and/or reconceptualization response shift, and 3) change due to change in the construct of interest. Subsequently, calculating distribution-based effect-size indices of change (i.e., standardized mean difference (SMD), standardized response mean (SRM), the probability benefit (PB), probability net befit (PNB), and the number needed to treat to benefit (NNTB)) enables evaluation and interpretation of the clinical significance of these different types of change. Change was investigated in health-related quality of life data from 170 cancer patients, assessed prior to surgery and three months following surgery. Results indicated that patients deteriorated on general physical health and general fitness (effect size SRM = -0.72, and SRM = -0.37), and improved on general mental health (SRM = 0.48). The decomposition of change showed that the impact of response shift on the assessment of change was small. We conclude that SEM can be used to enable the evaluation and interpretation of the impact of response shift effects on the assessment of change, particularly through calculation of effect-size indices of change. Insight into the occurrence and clinical significance of possible response shift effects will help to better understand changes in PROs.

Introduction

Patient-reported outcomes (PROs) have become increasingly important, both in clinical research and practice. PROs may include measures of subjective wellbeing, functional status, symptoms, or health-related quality of life (HRQL). The patient perspective on health provides insight into the effects of treatment and disease that is imperative for understanding health outcomes. PROs thus present important measures for evaluating the effectiveness of treatments and changes in disease trajectory, especially in chronic disease (Revicki, Hays, Cella, & Sloan, 2008), and palliative care (Ferrans, 2007).

The investigation and interpretation of change in PROs can be hampered because different types of change may occur. Differences in the scores of PROs are usually taken to indicate change in the construct that the PROs aim to measure. However, these differences can also occur because patients change the meaning of their self-evaluation. Sprangers & Schwartz (1999) proposed a theoretical model for change in the meaning of self-evaluations, referred to as 'response shift'. They distinguish three different types of response shift: recalibration refers to a change in respondents' internal standards of measurements; reprioritization refers to a change in respondents' values regarding the relative importance of subdomains; and reconceptualization refers to a change in the meaning of the target construct. As the occurrence of response shift may impact the assessment of change, the detection of possible response shift effects is important for the interpretation of change in PROs. One of the methods that can be used to investigate the occurrence of response shift is the structural equation modeling (SEM) approach (Oort, 2005). Advantages of the SEM approach are that it enables the operationalization and detection of the different types of response shift, and that it can be used to investigate change in the construct of interest (e.g., HRQL) while taking possible response shifts into account.

Although clinicians and researchers acknowledge the occurrence of response shift, little is known about the magnitude and clinical significance of those effects (Schwartz et al., 2006). The detection of response shift is usually guided by tests of statistical significance. Although statistical tests can be used to determine whether occurrences of response shift are *statistically significant*, they cannot be taken to imply that the result is also *clinically significant* (i.e., meaningful). Statistical significance tests protect us from interpreting effects as being 'real' when they could in fact result from random error fluctuations. However, statistical significance tests do not protect us from interpreting small, but trivial effects as being meaningful. Therefore, assessing the meaningfulness of change in PROs has been an important research focus (Cappelleri & Bushmakin, 2014; Sloan, Cella, & Hays, 2005), as it is imperative for translating results to patients, clinicians or health practitioners. However, there is no universally accepted approach to determine the meaningfulness of change in PROs (Wyrwich, et al., 2005).

One of the approaches that can be used to determine the clinical significance of change in PROs is to calculate distribution-based effect-size indices. Distribution-based effect sizes are calculated by comparing the change in outcome to a measure of the variability (e.g., a standard

deviation). The resulting effect sizes are thus standardized measures of the relative size of effects. They facilitate comparison of effects from different studies, particularly when outcomes are measured on unfamiliar or arbitrary scales (Coe, 2002). In addition, previous research has shown that distribution-based indices often lead to similar conclusions as when the clinical significance of effects is directly linked to patients' or clinicians' perspectives on the importance of change, i.e. so-called anchor-based indices of effects (Cella, et al., 2002; Eton, et al., 2004; Jayadevappa, Malkowicz, Wittink, Wein, & Chhatre, 2012). Furthermore, the interpretation of effect-size indices as indicating 'small', 'medium', or 'large' effects is possible using general 'rules of thumb' (e.g., Cohen, 1988). Therefore, distribution-based effect-size indices can be used to convey information about the clinical meaningfulness of results.

The aim of this paper is to explain the calculation of effect-size indices within the SEM framework for the investigation and interpretation of change. In addition, we explain how this enables the evaluation and interpretation of the impact of response shift on the assessment of change. Specifically, we use SEM to decompose observed change into change due to response shift, and change due to the construct of interest (i.e., 'true' change). Subsequently, we illustrate the calculation and interpretation of various effect-size indices, i.e. the standardized mean difference, the standardized response mean, the probability of benefit, the probability of net benefit, and the number needed to treat to benefit, for each component of the decomposition. This enables the evaluation of the contributions of response shift and true change to the overall assessment of change in the observed variables. To illustrate, we will use SEM to investigate change in data from 170 cancer patients, who's HRQL was assessed prior to surgery and three months following surgery. We aim to show that distribution-based effect-size indices can contribute to the clinical interpretability of change in PROs.

Method

Calculation of effect-size indices of change
Throughout this paper we will use the difference between the scores on a pre- and post-test as an example to explain the calculation of effect-size indices of change.

Standardized mean difference (SMD). One of the distribution-based methods to describe the magnitude of change, is to express the difference between pre- and post-test means in standard deviation units (see Table 1). The resulting standardized mean difference (SMD) can be estimated using sample statistics, where the observed sample means serve as estimates for the population means. However, for the estimation of the population standard deviation, there are various options. One option is to use the standard deviation of the pre-test. The resulting effect-size is a measure of change between pre- and post-test in terms of standardized units of between-subject variability at baseline (i.e., before the start of treatment). Other options for the calculation of the SMD effect size include using the pooled standard deviation (i.e., treating the

pre- and post-test assessments as if they were independent; Olejnik & Algina, 2000), and using the standard deviation of a subsample of patients (e.g., stable or improved patients; see Middel & van Sonderen, 2002). However, the pre-test standard deviation is thought to provide the best estimate of the population standard deviation as it is not yet affected by the occurrences between pre-and post-test (Kazis, Anderson, & Meenan, 1989). As this approach seems to be used most often in the literature (e.g., Copay, Subach, Glassman, Polly, & Schuler, 2007; Durlak, 2009; Hojat & Xu, 2004; Norman, Wyrwich, & Patrick, 2007; Schwartz, et al., 2006), we refer to the resulting effect size as the SMD effect size (see Table 1).

Standardized response mean (SRM). An alternative to using the pre-test standard deviation for the calculation of effect-size indices of change, is to use the standard deviation of the difference. In fact, this is what Cohen (1988) suggested as an appropriate effect-size index of change (p. 48), as it specifically takes into account the correlation between measurements (see Table 1). It provides a measure of change between pre- and post-test in terms of standardized units of between-subjects variability in change. The resulting effect size is known as the standardized response mean (SRM), which has been argued to be most intuitive and relevant for the interpretation of change (Liang, Fossel, & Larson, 1990). Moreover, using the standard deviation of the difference as a standardizer results in an estimate that is equivalent to a z-value, and thus facilitates the translation to other effect sizes (see Table 1). Therefore, in this paper we use the SRM effect size as the preferred effect-size index of change.

Table 1 | Calculation of effect-size indices of change

Definition	Calculation
Standardized mean difference (SMD)	
$\dfrac{\mu_{post} - \mu_{pre}}{\sigma}$	$\text{SMD} = \dfrac{\bar{X}_{post} - \bar{X}_{pre}}{sd_{pre}}$
Standardized response mean (SRM)	
$\dfrac{\mu_{post} - \mu_{pre}}{\sigma_{difference}}$	$\text{SRM} = \dfrac{\bar{X}_{post} - \bar{X}_{pre}}{\sqrt{sd_{post}^2 + sd_{pre}^2 - 2r_{post,pre}\, sd_{post} sd_{pre}}}$
Probability benefit (BP)	
$P(x_{post,i} > x_{pre,i})$	$\text{PB} = \Phi(\text{SRM})^{*}$
Probability net benefit (PNB)	
$P(x_{post,i} > x_{pre,i}) - P(x_{post,i} < x_{pre,i})$	$\text{PNB} = \Phi(\text{SRM}) - (1 - \Phi(\text{SRM})) = 2\Phi(\text{SRM}) - 1$
Number needed to treat to benefit (NNTB)	
$\dfrac{1}{\text{PNB}}$	$\text{NNTB} = \dfrac{1}{2\Phi(SRM) - 1}$

* where Φ is the cumulative standard normal distribution.

Interpretation of SMD and SRM effect sizes. As a general rule of thumb, values of 0.2, 0.5, and 0.8 of the SMD effect size can be interpreted as indicating 'small', 'medium', and 'large' effects respectively (Cohen, 1988). These rules of thumb were originally proposed for effect-size indices as calculated with the within population standard deviation, and can thus be considered appropriate for the interpretation of the SMD effect size of change as calculated with the pre-test standard deviation. However, it has been argued that application of these rules of thumb for the interpretation of the SRM effect size of change may lead to over- or under-estimation of effects (Middel & van Sonderen, 2002), as the relation between the SMD and SRM effect sizes depends on the correlation between the measurements (see Table 2). Specifically, the relation between the pre-test standard deviation (under the assumption that pre- and post-test population standard deviations are equal, i.e., $\sigma_{pre} = \sigma_{post} = \sigma$) and the standard deviation of the difference, can be formulated as follows: $\sigma_{difference} = \sigma \sqrt{2(1 - \rho_{pre,post})}$,[1] where $\rho_{pre,post}$ is the correlation between the pre- and post-test. Thus, interpretation of the SRM effect size of change according to the general rules of thumb may lead to an underestimation when the correlation between measurements is smaller than 0.5, and to an overestimation when the correlation between measurements is larger than 0.5. However, as it might not to be an unrealistic assumption that correlations between consecutive measurements are generally around 0.5, the rules of thumb for interpretation of the SRM effect size can be applied without a major risk of over- or under-valuation of the magnitude of effects.

Table 2 | The standardized response mean (SRM) a function of the standardized mean difference (SMD) and varying correlations between measurements

	Correlation between measurements										
SMD	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
0.20	0.14	0.15	0.16	0.17	0.18	0.20	0.22	0.26	0.32	0.45	∞
0.50	0.35	0.37	0.40	0.42	0.46	0.50	0.56	0.65	0.79	1.12	∞
0.80	0.57	0.60	0.63	0.68	0.73	0.80	0.89	1.03	1.26	1.79	∞

Relation to other effect-size indices of change

The SRM effect-size indices of change express the magnitude of change in terms of standard deviation units. To enhance clinical interpretability of the proposed effect-size index of change, we explain how this effect size can be converted into other well-known effect-size indices that have been proposed specifically for their intuitive (clinical) appeal.

Probability of benefit (PB). To enhance the interpretability of the magnitude of an effect, it has been proposed to use an estimate of the probability of a superior outcome (Grissom, 1994). In the context of pre- and post-test comparison this refers to the probability that a random subject shows a superior post-test score as compared to the pre-test score, i.e. the probability

---

[1] $\sqrt{\sigma_{post}^2 + \sigma_{pre}^2 - 2\rho_{pre,post}\, \sigma_{post}\sigma_{pre}} = \sqrt{\sigma^2 + \sigma^2 - 2\rho_{pre,post}\, \sigma^2} = \sqrt{2\sigma^2 - 2\rho_{pre,post}\, \sigma^2} = \sqrt{2\sigma^2(1 - \rho_{pre,post})} = \sigma \sqrt{2(1 - \rho_{pre,post})}$

that a random subject shows a positive change or improvement over time. We refer to this effect size as the probability of benefit (PB), but it is also known as the probability of superiority (PS; Grissom, 1994), the common language effect size (CLES; McGraw & Wong, 1992) and the area under the curve (AUC). The effect size was proposed specifically for its intuitive appeal and ease of interpretation, and has been recommended for developing insights about differences (Kraemer & Kupfer, 2006). The PB effect size is defined as a function of the SRM (see Table 1).

Probability of net benefit (BNP). The PB effect size does not take into account possible detrimental effects. That is, subjects may show a deterioration over time. The probability of such harmful effects (probability of harm; PH) is defined as 1 – PB. The effect size that we refer to as the probability of net benefit (PNB) is the difference between PB and PH (see Table 1). In the context of pre- and post-test comparison, the PNB is calculated as the difference between the probability that a random subject improves over time (i.e., PB), and the probability that a random subject deteriorates over time (i.e., PH). Or, in other words, the net probability that a random subject improves over time (i.e., PNB). This effect size is commonly applied to binary outcomes, where it is known as the success rate difference (SRD; Rosenthal & Rubin, 1982), absolute risk reduction (ARR), or risk difference (RD). It is one of the effect sizes that is recommended by the consolidated standards of reporting trials (Schulz et al., 2010).

Number needed to treat to benefit (NNTB). Another effect size that has been recommended for clinical interpretability (Kraemer & Kupfer, 2006) is the number needed to treat (NNT; Laupacis, Sackett, & Roberts, 1988). The NNT was originally defined for binary (success/ failure) outcomes, and refers to the number of subjects one would expect to treat to have one more success (or one less failure) as compared to the control group. It facilitates interpretation of effects in – clinically meaningful – terms of patients that need to be treated to reach a success rather than probabilities of a success (Sedgwick, 2015). The NNT for continuous outcomes is defined as the inverse of the PNB. As such, the NNT is also referred to as the number needed to treat to benefit (NNTB). In the context of pre- and post-test comparison, the NNTB can be interpreted as the expected number of patients that needs to be treated to have one more patient show an improvement (i.e., benefit) as compared to the expected number of patients who show a deterioration. When the PNB is negative (i.e., the net effect is harmful), then the NNTB is interpreted as the expected number of patients that needs to be treated to have one more patient show a deterioration (i.e., a harmful effect) as compared to the expected number of patients who show an improvement.

Relation between different effect-size indices of change. The relation between the SRM and the other effect-size indices of change (i.e., PB, PNB, and NNTB) can be used to derive the respective values of these effect-size indices that correspond to different values of the SRM effect size, including the 0.20, 0.50 and 0.80 thresholds for interpretation of 'small', 'medium' and

'large' effects respectively (see Table 3). For example, a medium SRM effect size corresponds to a PB effect size that indicates that 69% of patients show an improvement (i.e., PB = 0.69), where 38% more patients show an improvement as compared to a deterioration (PNB = 0.38), and three patients need to be treated to have one more patient show an improvement as compared to the number of expected patients who show a deterioration (NNTB = 2.61), i.e. for every three patients that are treated two patients will improve as compared to one patient who deteriorates.

Table 3 | Relation between different effect-size indices of change

SRM	PB	PNB	NNTB
0.00	0.50	0.00	∞
0.10	0.54	0.08	12.55
0.20	*0.58*	*0.16*	*6.31*
0.30	0.62	0.24	4.24
0.40	0.66	0.31	3.22
0.50	*0.69*	*0.38*	*2.61*
0.60	0.73	0.45	2.21
0.70	0.76	0.52	1.94
0.80	*0.79*	*0.58*	*1.74*
0.90	0.82	0.63	1.58
1.00	0.84	0.68	1.46
1.10	0.86	0.73	1.37
1.20	0.88	0.77	1.30
1.30	0.90	0.81	1.24
1.40	0.92	0.84	1.19
1.50	0.93	0.87	1.15
1.60	0.95	0.89	1.12
1.70	0.96	0.91	1.10
1.80	0.96	0.93	1.08
1.90	0.97	0.94	1.06
2.00	0.98	0.95	1.05

Notes: SRM = standardized response mean; PB = probability of benefit; PNB = probability of net benefit, NNTB = number needed to treat to benefit.

Decomposition of change

Within the SEM framework, the different response shifts are operationalized by differences in model parameters. Specifically, changes in the pattern of common factor loadings (Λ) are

indicative of reconceptualization response shift, changes in the values of the common factor loadings are indicative of reprioritization response shift, and changes in the intercepts (τ) are indicative of recalibration response shift. Changes in the means of the underlying factors (κ) are indicative of 'true' change, i.e., change in the underlying common factor. The contributions of the different types of response shifts and 'true' change to the changes in the observed variables can be investigated using the decomposition of change (see Table 4). Specifically, using the same standard deviation to standardize observed change, and the different elements of the decomposition, enables evaluation and interpretation of the contribution of recalibration, reprioritization and reconceptualization, and 'true' change, to the change in the observed variables. In addition, the overall impact of response shift on the assessment of change in the underlying construct of interest can be evaluated through the comparison of effect-size indices for change in the means of the underlying common factors before and after taking into account possible response shift effects.

Table 4 | Decomposition of observed change according to the SEM approach

Specification of observed means and observed change
Observed mean post-test
$\mu_{post} = \tau_{post} + \Lambda_{post}\kappa_{post}$
Observed mean pre-test
$\mu_{pre} = \tau_{pre} + \Lambda_{pre}\kappa_{pre}$
Observed change
$(\mu_{post} - \mu_{pre}) = (\tau_{post} + \Lambda_{post}\kappa_{post}) - (\tau_{pre} + \Lambda_{pre}\kappa_{pre})$
Decomposition of change
Observed change = Recalibration + Reprioritization & Reconceptualization + True change
$(\mu_{post} - \mu_{pre}) \quad = (\tau_{post} - \tau_{pre}) \quad + (\Lambda_{post} - \Lambda_{pre})\kappa_{post} \quad\quad + \Lambda_{pre}(\kappa_{post} - \kappa_{pre})$

Notes: The Greek symbols reflect the parameter estimates of observed factor means (μ), intercepts (τ), common factor loadings (Λ), and common factor means (κ).

Illustrative Example

To illustrate the calculation and interpretation of effect-size indices of change we used health-related quality of life (HRQL) data from 170 newly diagnosed cancer patients'. Patients' HRQL was assessed prior to surgery (pre-test) and three months following surgery (post-test). The sample included 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. These data have been used before to investigate response shift, and details about the study procedure, patient-characteristics and measurement instruments can be found elsewhere (Visser et al., 2013; Visser, Oort, & Sprangers, 2005; Verdam, Oort, Visser, & Sprangers, 2012).

Measures
HRQL was assessed using the SF-36 health survey (Ware et al., 1993) and the multidimensional fatigue inventory (Smets, Garssen, Bonke, & de Haes, 1995), resulting in the following nine 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), mental health (MH), and fatigue (FT). For computational convenience the scale scores were transformed so that they all ranged from 0 to 5, with higher scores indicating better health.

Measurement model
The measurement model is depicted 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.

The measurement model was the basis for a structural equation model for pre-, and post-test with no across measurement constraints. Imposition of equality constraints on all model parameters associated with response shift effects indicated the presence of response shift (see Verdam et al., 2012 for more information). Four cases of response shift were identified: reconceptualization of GH, reprioritization of SF as an indicator of GenPhys, and recalibration of RP and BP (see Figure 1).

Effect-size indices of change
The parameter estimates of the model in which all response shifts were taken into account were used for the decomposition of change to enable the calculation of effect-size indices of change, and the contributions to change of the different response shift effects and 'true' change (see Table 5).
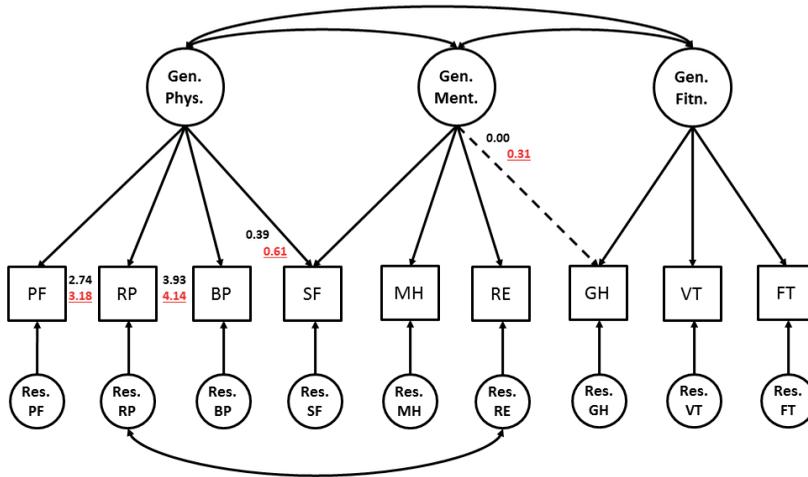
8

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 subscales of the HRQL questionnaires). Numbers are maximum likelihood estimates of the model parameters associated with response shift: common factor loadings (reprioritization and reconceptualization), and intercepts (recalibration). Values represent different pre-test (black) and post-test (red) estimates.

General Physical Health. There was an overall medium deterioration in GenPhys (standardized response mean (SRM) = -0.72). Conversion of this effect-size into probability of benefit (PB), the probability of net benefit (PNB), and the number needed to treat to benefit (NNTB) yielded values of 0.23, -0.53, and -1.88 respectively. This indicates that only 23% of patients showed an improvement over time (PB = 0.23), and that 53% more patients deteriorated than improved (PNB = -0.53). The NNTB indicates that with every 1.88 patients to be treated, there would be one more patient who shows a deterioration as compared to an improvement. In other words, two of every three patients who are treated are expected to show a deterioration.

The contribution of 'true' change (i.e., the change in the observed indicators that is due to change in the underlying common factors) was in the same direction and of similar magnitude for the indicators that load only on GenPhys (i.e., RF, RP and BP; see Table 5). The indicator SF loaded not only on GenPhys but also on GenMent, and therefore showed a deviating pattern of change. The contribution of 'true' change in this indicator was a combination of the deterioration of GenPhys and improvement of GenMent (see below), that cancelled each other out.

Three different response shifts were detected for the indicators of GenPhys. Patients' SF became more important to the measurement of GenPhys after treatment (with a contribution to change: SRM = -0.10). In addition, patients scored higher on RP and BP after treatment, as compared to the other indicators of GenPhys (with a contribution to change: SRM = 0.19, and SRM = 0.17 respectively). These occurrences of response shift thus had small effects on the change in the observed indicators. To illustrate, the response shift effect of BP can be translated

as follows (see Table 5): 57% of patients showed a relative improvement (PB = 0.57), with 14% more patients showing a relative improvement as compared to a relative deterioration (PNB = 0.14). For every seven patients who are treated there would be one more patient who shows a relative improvement due to recalibration response shift (NNTB = 7.21), i.e. four patients would show a relative improvement as compared to three patients who are expected to show a relative deterioration.

The influence of response shift on the assessment of change is apparent when we look at the estimated effect sizes for observed change. Here, we can see that the deterioration in RP and BP became somewhat smaller as was expected from only the change in GenPhys. In addition, the observed change in SF was slightly more negative, than what would be expected only from the changes in the underlying factors of GenPhys and GenMent. For the indicator PF there was no response shift detected, and thus the observed change was equal to the contribution of 'true' change (i.e., the observed change in the indicator could be ascribed to change in GenPhys). If response shift had not been taken into account, the change in the underlying common factor GenPhys would have been estimated to be slightly smaller (SRM = -0.59, instead of SRM = -0.72).

General Mental Health. There was an overall small improvement in GenMent (SRM = 0.48; PB = 0.69; PNB = 0.37; NNTB = 2.69). The contribution of 'true' change in the indicators that load only on GenMent (MH and RE) was in the same direction and of similar magnitude (see Table 5). There were no response shifts detected for these indicators, and thus all observed change could be described to 'true' change.

Reconceptualization was detected for the indicator GH, which became indicative of GenMent after treatment. The contribution of 'true' change in the decomposition of change for GH showed a small deterioration (SRM = -0.15), reflecting not only the contribution of 'true' change (deterioration) in GenFitn (see below), but also the contribution of 'true' change (improvement) in GenMent. The observed change in GH was thus less negative than what would be expected only due to 'true' change in GenFitn. This contribution of reconceptualization response shift of GH (with a contribution to change of SRM = .14) explains the deviating pattern of observed change in de indicator GH (SRM = -.01). Although the detected response shift had a small impact on the assessment of change at the level of the indicator, it did not influence the overall change in the underlying common factor GenMent. If response shifts had not been taken into account, the change in GenMent would have been estimated to be of similar magnitude (SRM = 0.45 instead of SRM = 0.48).

General Fitness. There was an overall small deterioration of GenFitn (SRM = -0.37; PB = 0.35; PNB = -0.29; NNTB = -3.44). The two indicators (VT and FT) that loaded only on GenFitn showed a deterioration in the same direction and with similar magnitude. There was no response shift detected for these indicators, and thus the observed change in these indicators could be attributed to 'true' change.

8

Table 5 | Effect-size indices of (contributions to) change for the decomposition of change

Scale	Cohen's SRM	PB	PNB	NNTB
Observed change: $(\mu_{post} - \mu_{pre})$				
PF	-0.51	0.30	-0.39	-2.54
RP	-0.28	0.39	-0.22	-4.61
BP	-0.25	0.40	-0.19	-5.16
SF	-0.09	0.46	-0.07	-13.73
MH	0.37	0.64	0.29	3.49
RE	0.26	0.60	0.21	4.85
GH	-0.01	0.49	-0.01	-97.85
VT	-0.31	0.38	-0.25	-4.06
FT	-0.32	0.37	-0.25	-3.94
Response shift: $(\tau_{post} - \tau_{pre}) + (\varLambda_{post} - \varLambda_{pre})\varkappa_{post}$				
PF	-	-	-	-
RP	0.19[a]	0.58	0.15	6.51
BP	0.17[a]	0.57	0.14	7.21
SF	-0.10[b]	0.46	-0.08	-12.56
MH	-	-	-	-
RE	-	-	-	-
GH	0.14[c]	0.55	0.11	9.14
VT	-	-	-	-
FT	-	-	-	-
True Change: $\varLambda_{pre}(\varkappa_{post} - \varkappa_{pre})$				
PF	-0.51	0.30	-0.39	-2.54
RP	-0.47	0.32	-0.36	-2.77
BP	-0.42	0.34	-0.33	-3.07
SF	0.01	0.50	0.01	159.57
MH	0.37	0.64	0.29	3.49
RE	0.26	0.60	0.21	4.85
GH	-0.15	0.44	-0.12	-8.36
VT	-0.31	0.38	-0.25	-4.06
FT	-0.32	0.37	-0.25	-3.94

Notes: N = 170; SRM = standardized response mean, where values of 0.2, 0.5, and 0.8 indicate small, medium, and large effects; PB = probability of benefit; PNB = probability of net benefit, NNTB = number needed to treat to benefit. [a] = recalibration, [b] = reprioritization, [c] = reconceptualization.

Discussion

In this paper we have shown how to calculate effect-size indices of change using structural equation modeling (SEM). We used SEM for the decomposition of change, where observed change (e.g., change in the subscales of a health-related quality of life (HRQL) questionnaire) is decomposed into change due to recalibration, reprioritization and reconceptualization, and 'true' change in the underlying construct (e.g., HRQL). Calculation of effect-size indices for each of the different elements of the decomposition enables the evaluation and interpretation of the impact of response shift on the assessment of change.

We used distribution-based effect sizes to interpret and evaluate the magnitude of change, and the impact of response shift on the assessment of change. Specifically, we proposed to use the standardized response mean as the preferred effect size of change. Results from our illustrative example indicated that patients experienced small to medium sized changes in their scores on the subscales of the HRQL questionnaires. Four response shifts were detected, but the impact of the detected response shift on the assessment of change was small; both at the level of the observed variables and at the level of the underlying common factors. Similar sizes of effects were reported in a meta-analysis on response shift (Schwartz et al., 2006), although these results were based on studies that did not use SEM methodology. Moreover, the authors concluded that a lack in standards on reporting effect-size indices prevented definitive conclusion on the clinical significance of response shift. The decomposition of change and the proposed calculation of effect-size indices may advance the standard reporting and comparison of results, and thus facilitate the interpretation and impact of the different types of change in PROs. This may help to translate the findings of response-shift research into something that is tangible to patients, clinicians and researchers alike.

Some limitations of distribution-based effect sizes should be noted. Distribution-based indices may be influenced by the reliability of the measurement, as unreliable measurement will result in larger standard deviations and thus smaller effect sizes. In addition, when the assumption of normal distributions is not tenable this may alter the interpretation of the effect size, which hinders the comparison of effect-size indices from different samples or studies. Finally, restriction of range has also been mentioned as a limitation of distribution-based indices. However, the fact that the clinical significance of an effect is calculated and interpreted relative to the variation within a sample could also be considered a strength. For example, it may be difficult to define the absolute change that indicates clinical significance, as smaller changes in one group of patients may be more meaningful than larger changes in another group of patients. The effect size of change is calculated using the variability of change within a patient group, and will thus provide an interpretation of the relative – instead of absolute – importance of the effect. Nevertheless, one should take into consideration the context of the study when interpreting the magnitude of the effects. Keeping the general limitations of distribution-based indices in mind, it is recommended that the proposed effect size of change is used as a guideline

for the interpretation of clinical significance, rather than a rule (Guyatt et al., 2002). Conversion of the effect size of change to the probability of benefit, the probability of net benefit, or the number needed to treat to benefit may enhance clinical interpretability of effects. Different effect sizes, or indices of clinical significance in general, can complement each other as they facilitate different tasks and insights.

SEM provides a valuable tool for the assessment of change, and investigation of response shift in PROs. The decomposition of change – and subsequent calculation of effect-size indices – provides insight into the impact of response shifts on the assessment of change. Advancing the standard reporting of effect-size indices of change will enhance the comparison of effects, facilitate future meta-analysis, and provides insight into the size of the effects instead of merely their statistical significance. As such, the use of effect-size indices of change can facilitate progress in our endeavors of evaluating and interpreting clinical significant changes in PROs.