Structural equation modeling–based effect-size indices were used to evaluate and interpret the impact of response shift effects

Verdam, M.G.E.; Oort, F.J.; Sprangers, M.A.G.

DOI
10.1016/j.jclinepi.2017.02.012

Publication date
2017

Document Version
Final published version

Published in
Journal of Clinical Epidemiology

Journal of Clinical Epidemiology 85 (2017) 37–44

Journal of Clinical Epidemiology

Structural equation modeling–based effect-size indices were used to evaluate and interpret the impact of response shift effects

Mathilde G.E. Verdam[a,b,*], Frans J. Oort[b], Mirjam A.G. Sprangers[a]

[a]*Department of Medical Psychology, Academic Medical Centre, University of Amsterdam, Meibergdreef 15, Amsterdam, 1105 AZ, The Netherlands*
[b]*Department of Child Development and Education, University of Amsterdam, Postbus 15776, 1001 NG, Amsterdam, The Netherlands*

Accepted 21 February 2017; Published online 22 March 2017

Abstract

Objectives: The 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.

Study Design and Setting: When response shift is investigated through structural equation modeling (SEM), observed change can be decomposed into change because of recalibration response shift, change because of reprioritization and/or reconceptualization response shift, and change because of change in the construct of interest. Subsequently, calculating effect-size indices of change enables evaluation and interpretation of the clinical significance of these different types of change.

Results: Change was investigated in health-related quality of life data from 170 cancer patients, assessed before surgery and 3 months after surgery. Results indicated that patients deteriorated on general physical health and general fitness and improved on general mental health. The decomposition of change showed that the impact of response shift on the assessment of change was small.

Conclusion: 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. Insight into the occurrence and clinical significance of possible response shift effects will help to better understand changes in PROs.

Keywords: Effect size; Clinical significance; Patient-reported outcomes; Health-related quality of life; Structural equation modeling; Change assessment

1. Introduction

Patient-reported outcomes (PROs) have become increasingly important, both in clinical research and practice. PROs may include measures of subjective well-being, 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 [1] and palliative care [2].

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 and Schwartz [3] proposed a theoretical model for change in the meaning of self-evaluations, referred to as a 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. To illustrate, when a patient is being asked to fill in a questionnaire about quality of life, he or she may indicate to be limited in social functioning (SF) very often before treatment and some of the time after treatment. The change in these responses can be interpreted as a reduction in SF and indicative of a reduction in quality of life. However, the observed change may also occur because the patient has recalibrated what very often means, for example, the response very often may refer to many more times after treatment than it did before treatment. With reprioritization response shift,

Conflict of interest: The authors have no conflict of interest to declare.

* Corresponding author. Tel.: +31205251359; fax: +31205251500.
E-mail address: m.g.e.verdam@uva.nl (M.G.E. Verdam).

http://dx.doi.org/10.1016/j.jclinepi.2017.02.012
0895-4356/

What is new?

Key findings

In this article, we explain how the impact of response shift on the assessment of change can be evaluated through the calculation of effect-size indices of change to convey information about the clinical meaningfulness of results.

What this adds to what was known?

Structural equation modeling provides a valuable tool for the assessment of change, and investigation of response shift in patient-reported outcomes (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.

What is the implication and what should change now?

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. Insight into the occurrence and clinical significance of possible response shift effects will help to better understand changes in PROs.

the observed change may occur because the relative importance of SF to the patient's quality of life increased. Finally, with reconceptualization response shift, the meaning of a patient's response may have changed, for example, a patient may interpret SF as work related before treatment and as family related after treatment.

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 Oort's structural equation modeling (SEM) approach [4]. 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. We note, however, that SEM is a group level analysis and will only detect response shifts that affect a substantial part of the sample.

Although clinicians and researchers acknowledge the occurrence of response shift, little is known about the magnitude and clinical significance of those effects [5]. 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 [6,7] 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 [8].

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 [9]. 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, that is, the so called anchor-based indices of effects [10–12]. Furthermore, the interpretation of effect-size indices as indicating small, medium, or large effects is possible using general rules of thumb (e.g., Ref. [13]). Therefore, distribution-based effect-size indices can be used to convey information about the clinical meaningfulness of results.

The aim of this article 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 because of response shift and change because of the construct of interest (i.e., true change). Subsequently, we illustrate the calculation and interpretation of various effect-size indices, that is, the standardized mean difference (SMD), standardized response mean (SRM), probability of benefit (PB), probability of net benefit (PNB), and number needed to treat to benefit (NNTB), 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, whose HRQL was assessed before surgery and 3 months after surgery. We aim to show that distribution-based effect-size indices can contribute to the clinical interpretability of change in PROs.

2. Method

2.1. Calculation of effect-size indices of change

Later we explain the calculation and interpretation of different effect-size indices of change using pretest and post-test comparison as an example. A more detailed explanation of the (statistical) derivations of these effect-size indices and their inter-relationships is offered in an online Technical Appendix (see on the journal's Web site at www.elsevier.com).

2.1.1. Standardized mean difference

One of the distribution-based methods to describe the magnitude of change is to express the difference between pretest and post-test means in standard deviation units. The resulting SMD can be calculated using the standard deviation of the pretest (Table 1). This effect size thus expresses the magnitude of change in terms of variability between subjects at baseline, that is, before the start of treatment. The advantage of using the pretest standard deviation as a standardizer is that it is not yet affected by the occurrences between pretest and post-test [14]. Although other options for the calculation of the SMD effect size exist (see A.1 of the Technical Appendix for more details), the pretest standard deviation seems to be used most often in the literature (e.g., Refs. [5,15–18]). Therefore, we refer to the resulting effect size as the SMD effect size (Table 1).

2.1.2. Standardized response mean

An alternative to using the pretest 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 [13] suggested as an appropriate effect-size index of change (p. 48) as it specifically takes into account the correlation between pretest and post-test assessments (Table 1). The resulting effect size is known as the SRM. It expresses the magnitude of change in terms of between-subject variability in change, which has been argued to be most intuitive and relevant for the interpretation of effects [19]. 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 (Table 1). Therefore, in this article, we use the SRM effect size as the preferred effect-size index of change.

Table 1. Calculation of effect-size indices of change

Effect size	Calculation
SMD	$\frac{\overline{X}_{post}-\overline{X}_{pre}}{SD_{pre}}$
SRM	$\frac{\overline{X}_{post}-\overline{X}_{pre}}{\sqrt{SD^2_{post}+SD^2_{pre}-2r_{post,pre}\,SD_{post}SD_{pre}}}$
PB	$\Phi(\text{SRM})$[a]
PNB	$\Phi(\text{SRM}) - (1 - \Phi(\text{SRM})) = 2\Phi(\text{SRM}) - 1$
NNTB	$\frac{1}{2\Phi(\text{SRM}) - 1}$

Abbreviations: SMD, standardized mean difference; SD, standard deviation; SRM, standardized response mean; PB, probability benefit; PNB, probability net benefit; NNTB, number needed to treat to benefit.

[a] Φ is the cumulative standard normal distribution.

2.1.3. 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 [13]. It has been argued that application of these rules of thumb for the interpretation of the SRM effect size of change may lead to overestimation or underestimation of effects [20]. Specifically, the interpretation of the SRM effect size of change according to the general rules of thumb may lead to an underestimation when the correlation between pretest and post-test measurements is smaller than 0.5 and to an overestimation when the correlation between measurements is larger than 0.5 (see A.2 of the Technical Appendix for more details). However, as it might not 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 overvaluation or undervaluation of the magnitude of effects.

2.2. 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.

2.2.1. Probability of benefit

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 [21]. In the context of pretest and post-test comparison, this refers to the probability that a random subject shows a superior post-test score as compared with the pretest score, that is, the probability that a random subject shows a positive change or improvement over time. We refer to this effect size as the PB, but it is also known as the probability of superiority [21], the common language effect size [22], and the area under the curve. The effect size was proposed specifically for its intuitive appeal and ease of interpretation and has been recommended for developing insights about differences [23]. The PB effect size can be calculated using the SRM effect size (Table 1).

2.2.2. Probability of net benefit

The PB effect size does not take into account possible detrimental effects. That is, subjects may show a deterioration over time. The effect size that we refer to as the PNB

(Table 1) is 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., 1 − PB). Or, in other words, the net probability that a random subject improves over time. This effect size is commonly applied to binary outcomes (e.g., success/failure), where it is known as the success rate difference [24], absolute risk reduction, or risk difference. It is one of the effect sizes that is recommended by the consolidated standards of reporting trials [25].

2.2.3. Number needed to treat to benefit

Another effect size that has been recommended for clinical interpretability [23] is the number needed to treat (NNT [26]). In the context of pretest and post-test comparison, the NNT can be interpreted as the expected number of patients who needs to be treated to have one more patient who show an improvement (i.e., benefit) as compared with the expected number of patients who show a deterioration. We refer to this effect size as the NNTB (Table 1) as it is calculated by taking the inverse of the PNB effect size. It facilitates interpretation of effects in—clinically meaningful—terms of patients who need to be treated to reach a success rather than probabilities of a success [27]. When the net effect is negative (i.e., more patients show a deterioration as compared with an improvement), then the NNTB is interpreted as the expected number of patients who needs to be treated to have one more patient who show a deterioration as compared with an improvement.

2.2.4. 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 A.3 of the Technical Appendix). 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 with a deterioration (PNB = 0.38), and three patients need to be treated to have one more patient who show an improvement as compared with the number of expected patients who show a deterioration (NNTB = 2.61), that is, for every three patients who are treated, two patients will improve as compared with one patient who deteriorates.

2.3. Decomposition of change

Within the SEM approach as proposed by Oort [4], the different response shifts are operationalized by differences in model parameters. Specifically, changes in the pattern of common factor loadings 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 are indicative of recalibration response shift.[1] Changes in the means of the underlying factors are indicative of true change, that is, 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 (i.e., observed questionnaire scores) can be investigated using the decomposition of change (see A.4 of the Technical Appendix for more details). Subsequently, 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.

3. Illustrative example

To illustrate the calculation and interpretation of effect-size indices of change, we used HRQL data from 170 newly diagnosed cancer patients. Patients' HRQL was assessed before surgery (pretest) and 3 months after 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 [28–30].

3.1. Measures

HRQL was assessed using the 36-Item Short Form Health Survey [31] and the multidimensional fatigue inventory [32], resulting in the following nine scales: physical functioning (PF), role limitations because of physical health (role—physical, RP), bodily pain (BP), general health perceptions (GH), vitality (VT), SF, role limitations because of emotional problems (role—emotional, RE), mental health (MH), and fatigue (FT). For computational

[1] Oort [4] distinguishes between uniform and nonuniform recalibration response shifts, where uniform recalibration is indicated when there are differences in intercept and nonuniform recalibration is indicated when there are differences in residual factor variances. As residual factor variances do not feature in the mean structure, they are not important for the assessment of change and not considered in this article.

convenience, the scale scores were transformed so that they all ranged from 0 to 5, with higher scores indicating better health.

3.2. Measurement model

The measurement model is depicted in Fig. 1 (see Ref. [33] 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, and others. The residual factors represent all that is specific to PF, RP, BP, and others, plus random error variation.

The measurement model was the basis for a structural equation model for pretest 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 Ref. [28] 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 (Fig. 1).

3.3. 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 (Table 2).

3.3.1. General physical health

There was an overall medium deterioration in GenPhys (SRM = −0.72). Conversion of this effect size into PB, PNB, and 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 with 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 because of 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; Table 2). The SF indicator loaded not only on GenPhys but also on GenMent, and therefore, it showed a deviating pattern of change. The

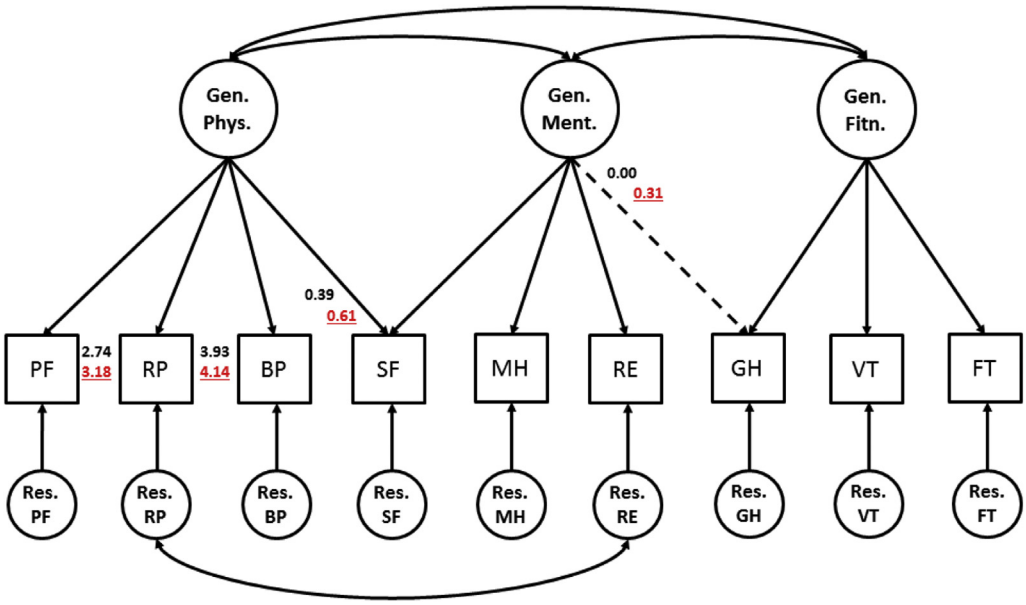

Fig. 1. The measurement model used in response shift detection. Circles represent latent variables (common and residual factors), and squares represent observed variables (the subscales of the health-related quality of life questionnaires). Numbers are maximum likelihood estimates of the model parameters associated with response shift: common factor loadings (reprioritization and reconceptualization) and intercepts (recalibration). Values represent different pretest (black) and post-test (red) estimates. These values are taken from the study by Verdam et al. [28] and differ slightly from the results of Oort et al. [33] because the former study also included the then-test assessment in the model. PF, physical functioning; RP, role—physical; BP, bodily pain; SF, social functioning; MH, mental health; RE, role—emotional; GH, general health perceptions; VT, vitality; FT, fatigue. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 2. Effect-size indices of (contributions to) change for the decomposition of change

Scale	SRM	PB	PNB	NNTB
Observed change				
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				
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				
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

Abbreviations: 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; PF, physical functioning; RP, role—physical; BP, bodily pain; SF, social functioning; MH, mental health; RE, role—emotional; GH, general health perceptions; VT, vitality; FT, fatigue.

$n = 170$.

[a] Recalibration.

[b] Reprioritization.

[c] Reconceptualization.

contribution of true change in this indicator was a combination of the deterioration of GenPhys and improvement of GenMent (see later) that canceled 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 with 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 (Table 2): 57% of patients showed a relative improvement (PB = 0.57), with 14% more patients showing a relative improvement as compared with a relative deterioration (PNB = 0.14). For every seven patients who are treated, there would be one more patient who shows a relative improvement because of recalibration response shift (NNTB = 7.21), that is, four patients would show a relative improvement as compared with 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 PF indicator, 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).

3.3.2. 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 (Table 2). There were no response shifts detected for these indicators, and thus all observed changes could be described to true changes.

Reconceptualization was detected for the GH indicator, 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 later) 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 because of true change in GenFitn. This contribution of reconceptualization response shift of GH (with a contribution to change of SRM = 0.14) explains the deviating pattern of observed change in the GH indicator (SRM = −0.01). 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).

3.3.3. 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.

4. Discussion

In this article, we have shown how to calculate effect-size indices of change using SEM. We used SEM for the decomposition of change, where observed change (e.g., change in the subscales of a HRQL questionnaire) is decomposed into change because of 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 SRM 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 [5], 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 [34]. Conversion of the effect size of change to the PB, PNB, or NNTB 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. Although they are all expressions of the same magnitude of effects—and thus different effect sizes may not necessarily convey new information—some can be more (clinically) intuitive and/or relevant for translating the meaningfulness of effects in certain contexts than others. For example, the probability benefit may be most informative when improvement is the most important outcome. However, when possible deterioration is important to consider, then the probability net benefit may be more informative. The NNTB may be especially useful when the cost of treatment is the focus of the study. Finally, the proposed effect size of change, the SRM, can be used to derive all these effect-size measures and may thus be the most informative index for comparison of results. Thus, each effect size has its own merits and advantages, depending on the context and purpose of the study, in facilitating the interpretation of clinical meaningfulness of results.

The present study focused on the investigation of change at the group level. One should keep in mind that group-level change is not the same as individual-level change, for example, some patients may show no change or even negative change. The same is true for changes because of response shift effects. The SEM method will only detect group-level response shift when most patients show individual-level response shift [4]. When most patients experience a response shift, this can be meaningful for the interpretation of general patterns of change—although some patients may not experience the detected response shift or not all patients experience the detected response shift to the same degree. Further investigation of such findings may show which (subgroup of) patients are prone to show the detected response shift and may help to understand why certain patients do or do not experience the response shift. Information about change and response shifts within groups of patients may thus eventually also enhance our understanding of individual experiences.

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.

Acknowledgments

This research was supported by the Dutch Cancer Society (KWF grant 2011-4985). F.J. Oort and M.G.E. Verdam participated in the Research Priority Area Yield of the University of Amsterdam. We thank M.R.M. Visser from the Academic Medical Center of the University of Amsterdam for making the data that was used in this study available for secondary analysis.

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jclinepi.2017.02.012.

References

[1] Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol 2008;61:102−9.
[2] Ferrans CE. Differences in what quality-of-life instruments measure. J Natl Cancer Inst Monogr 2007;37:22−6.
[3] Sprangers MAG, Schwartz CE. Integrating response shift into health-related quality of life research: a theoretical model. Soc Sci Med 1999;48:1507−15.
[4] Oort FJ. Using structural equation modeling to detect response shifts and true change. Qual Life Res 2005;14:587−98.
[5] Schwartz CE, Bode R, Repucci N, Becker J, Sprangers MAG, Fayers PM. The clinical significance of adaptation to changing health: a meta-analysis of response shift. Qual Life Res 2006;15:1533−50.
[6] Cappelleri JC, Bushmakin AG. Interpretation of patient-reported outcomes. Stat Methods Med Res 2014;23:460−83.
[7] Sloan JA, Cella D, Hays RD. Clinical significance of patient-reported data: another step toward consensus. J Clin Epidemiol 2005;58:1217−9.
[8] Wyrwich KW, Bullinger M, Aaronson N, Hays RD, Patrick DL, Symonds T, The Clinical Significance Consensus Meeting Group. Estimating clinically significant differences in quality of life outcomes. Qual Life Res 2005;14:285−95.
[9] Coe, R. (2002). It's the effect size, stupid: what effect size is and why it is important. Paper presented at the Annual Conference of the British Educational Research Association, University of Exeter, Exeter, England. Available at: www.leeds.ac.uk/educol/documents/00002182.htm. Accessed January 18, 2017.
[10] Cella D, Eton DT, Fairclough DL, Bonomi P, Heyes AE, Silberman C, et al. What is clinically meaningful change on the Functional Assessment of Cancer Therapy-Lung (FACT-L) Questionnaire? Results from Eastern Cooperative Oncology Group (ECOG) Study 5592. J Clin Epidemiol 2002;55:285−95.
[11] Eton DT, Cella D, Yost KJ, Yount SE, Peterman AH, Neuberg DS, et al. A combination of distribution- and anchor-based approaches determined minimally important differences (MIDs) for four endpoints in a breast cancer scale. J Clin Epidemiol 2004;57:898−910.
[12] Jayadevappa R, Malkowicz SB, Wittink M, Wein AJ, Chhatre S. Comparison of distribution- and anchor-based approaches to infer change in health-related quality of life of prostate cancer survivors. Health Res Educ Trust 2012;47:1902−25.
[13] Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale, NJ: Erlbaum; 1988.
[14] Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. Med Care 1989;27:S178−89.
[15] Copay AG, Subach BR, Glassman SD, Polly DW, Schuler T. Understanding the minimum clinically important difference: a review of concepts and methods. Spine J 2007;7:541−6.
[16] Durlak JA. How to select, calculate, and interpret effect sizes. J Pediatr Psychol 2009;34:917−28.
[17] Hojat M, Xu G. A visitor's guide to effect sizes: statistical significance versus practical (clinical) importance of research findings. Adv Health Sci Educ Theory Pract 2004;9:241−9.
[18] Norman GR, Wyrwich KW, Patrick DL. The mathematical relationship among different forms of responsiveness coefficients. Qual Life Res 2007;16:815−22.
[19] Liang MH, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopedic evaluation. Med Care 1990;28:632−42.
[20] Middel E, van Sonderen E. Statistical significant change versus relevant or important change in (quasi) experimental design: some conceptual and methodological problems in estimating magnitude of intervention-related change in health service research. Int J Integr Care 2002;2:1−18.
[21] Grissom RJ. Probability of the superior outcome of one treatment over another. J Appl Psychol 1994;79:314−6.
[22] McGraw KO, Wong SP. A common language effect size statistic. Psychol Bull 1992;111:361−5.
[23] Kraemer HC, Kupfer DJ. Size of treatment effects and their importance to clinical research and practice. Biol Psychiatry 2006;59:990−6.
[24] Rosenthal R, Rubin DB. A simple, general purpose display of magnitude of experimental effect. J Educ Psychol 1982;74:166−9.
[25] Schulz KF, Altman DG, Moher D, the CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. Ann Intern Med 2010;152:726−32.
[26] Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. N Engl J Med 1988;318:1728−33.
[27] Sedgwick P. Measuring the benefit of treatment: number needed to treat. Br Med J 2015;350:h2206.
[28] Verdam MGE, Oort FJ, Visser MRM, Sprangers MAG. Response shift detection through then-test and structural equation modeling: decomposing observed change and testing tacit assumptions. Neth J Psychol 2012;67:58−67.
[29] Visser MRM, Oort FJ, Sprangers MAG. Methods to detect response shift in quality of life data: a convergent validity study. Qual Life Res 2005;14:629−39.
[30] Visser MRM, Oort FJ, van Lanschot JJB, van der Velden J, Kloek JJ, Gouma DJ, et al. The role of recalibration response shift in explaining bodily pain in cancer patients undergoing invasive surgery: an empirical investigation of the Sprangers and Schwartz model. Psychooncology 2013;22:515−22.
[31] Ware JE, Snow KK, Kosinski M, Gandek B. SF-36 health survey: manual and interpretation guide. Boston, MA: The Health Institute, New England Medical Center; 1993.
[32] Smets EMA, Garssen B, Bonke B, De Haes JCJM. The Multidimensional Fatigue Inventory (MFI): psychometric qualities of an instrument to assess fatigue. J Psychosom Res 1995;39:315−25.
[33] Oort FJ, Visser MRM, Sprangers MAG. An application of structural equation modeling to detect response shifts and true change in quality of life data from cancer patients undergoing invasive surgery. Qual Life Res 2005;14:599−609.
[34] Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR, The Clinical Significance Consensus Meeting Group. Methods to explain the clinical significance of health status measures. Mayo Clinic Proc 2002;77:371−83.