



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].

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

CHAPTER 1

Introduction and Thesis Outline

Investigating Change in Patient-Reported Outcomes

Patient-reported outcomes (PROs) are increasingly recognized as a critical endpoint in health care and medicine, and routine assessment of PROs is becoming standard part of clinical practice (Fayers & Machin, 2016). PROs are used to provide a comprehensive picture of patients' wellbeing that can complement and broaden the information that is obtained from clinical outcomes, such as survival or disease progression. Implementing PROs into clinical practice helps to understand the impact of illness from the patient's viewpoint, and can make an important contribution to health care evaluations (Allison, Locker, & Feine, 1997).

The importance of measuring PROs, such as health-related quality of life (HRQL), is especially salient in view of ageing societies and more powerful health care interventions, which have led to an increasing number of people living with chronic disease (Huber et al., 2011). When disease cannot be cured, its treatment will target patients' wellbeing instead. In addition, treatment of fatal diseases, such as cancer, often result in minimal gain of survival, while therapeutic interventions come with serious side effects and impairments that heavily impact patients' lives (Osoba, 2011). In these situations, the wider effectiveness of treatment in terms of overall quality of life is thus of greater relevance. In clinical trials that focus on palliation or the end of life, patients' wellbeing may even become of primary importance. That is, the ultimate purpose of health care interventions may not be prolonged survival, but maintenance or optimization of patients' quality of life (Ferrans, 2007). HRQL assessments provide the opportunity to elucidate the effects of disease and treatment on patients' lives (Giesler & Williams, 1998), and make important contributions to the interpretations and conclusions of clinical trials (Barofsky, 2012; Editorial The Lancet, 1995; Fayers & Machin, 2016; Ferrans, 2007; Osoba, 1999, 2011).

Evaluating the impact of disease and treatment on patients' HRQL requires longitudinal assessment: the assessment of change provides insight into patients' perceived health trajectories. However, interpretation of change in HRQL outcomes is complicated by the fact that it is measured through self-assessment. Self-assessment brings about the problem that respondents may change their frames of references when answering HRQL-items. That is, people with similar quality of life over time may nevertheless score differently on HRQL-measures because the meaning of their self-evaluations has changed. Sprangers & Schwartz (1999) proposed a theoretical model for change in the meaning of self-evaluations, which they called 'response shift', a term coined by Howard et al. (1979). Sprangers & Schwartz distinguish three different types of response shift: *recalibration* refers to a change in respondents' internal criteria with which they assess the construct of interest; *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. The assessment of change in HRQL outcomes is thus hindered by the fact that different types of change may occur.

Change in the observed scores of a HRQL questionnaire may be due to change in the experienced quality of life, or – partly – due to change in the meaning of patients' self-evaluation. For example, a patient may indicate to experience nausea 'very often' before treatment, and 'some of the time' after treatment. The change in these responses can be interpreted as a reduction in nausea, and indicative of an improvement of HRQL. However, there might be other reasons for this change. It may occur because the patient has *recalibrated* what 'very often' means. When the experience of nausea occurs frequently due to treatment, the response 'very often' may refer to many more times after treatment than it did before treatment. Or, maybe the patient has adapted to the experience of nausea and therefore does not find the experience so overwhelming anymore – even though the patient is nauseated just as often. These are examples of recalibration response shift, as the internal criteria to assess the quality-of-life item have changed. With *reprioritization* response shift, there is a change in the relative importance of subdomains or items of HRQL. For example, when a disease or treatment causes serious physical impairment, social functioning may become more important to the patient's HRQL than physical functioning. As a result, observed change in social functioning does not (only) reflect change in HRQL, but also change in the importance of the subdomain. Finally, *reconceptualization* response shift refers to a change in the meaning of a patient's response, e.g., a patient may interpret 'social functioning' as work-related before treatment, and as family-related after treatment. Change in the indicator social functioning may in this case (at least partly) reflect a change in the meaning of the indicator itself. Thus, the investigation of response shift is important, as the occurrence of response shift may impact the assessment of change.

In this thesis, change in HRQL outcomes is investigated using structural equation modeling (SEM). There are several methodological approaches for the investigation of response shift, i.e. individualized methods, preference-based methods, successive comparison methods, design approaches, qualitative approaches, and statistical approaches (Schwartz & Sprangers, 1999). SEM is one of the statistical approaches, and has many advantages, including the possibility to distinguish between different types of change in the meaning of one's self-evaluation (i.e., response shifts), and the possibility to take into account possible response shift effects in order to enable a more valid assessment of change in HRQL.

Conceptualization and Operationalization of Health-related Quality of Life

Medical decision making research has focused increasingly on HRQL as an important outcome (Crosby, Kolotkin, & Williams, 2003), and many clinical trial organizations require or recommend the assessment of HRQL as a standard part of new trials (Editorial The Lancet, 1995; Fayers & Machin, 2016). But how is HRQL assessed, and how are these measures interpreted? Even though the assessment of HRQL is supported by, and readily applied in clinical practice and research, it is also generally acknowledged that there is a lack of theoretical agreement on the concept of HRQL (e.g., Aaronson, 1991; Barofsky, 2012; Fayers & Machin, 2016; Ferrans, 2007; Gill & Feinstein, 1994; Hunt, 1997). This has led to heterogeneity in both

its conceptualization and operationalization, which complicates the interpretation of HRQL outcomes. Moreover, heterogeneity in conceptualization and operationalization of HRQL may also influence the occurrence and interpretation of different types of change.

Heterogeneity in conceptualization of HRQL. The interest in the patient-perspective on health was first encouraged by the redefinition of health in 1948 by the World Health Organization (WHO) as “not only the absence of disease and infirmity, but also the presence of physical, mental, and social wellbeing”.¹ This redefinition of health formalized the idea that health outcomes should not only encompass clinical diagnostics such as the presence, or the number and size of tumors, but should also include the patients’ perspective on their wellbeing. Earlier developments within the medical setting focused on objective measurements of functional health status (e.g., the ability to perform routine tasks). Due to the WHO redefinition of health, these measures were broadened to indicate ‘quality of survival’ (e.g., symptomatic effects of treatment like pain, distress or suffering). The WHO redefinition of health also triggered an expansion of measures to evaluate the general wellbeing of populations that included person’s perceptions of life (e.g., social and emotional aspects). These developments have led to assessment scales that combine medical-health status indexes (i.e., physical and functional status) and social-science population indexes (i.e., social and psychological status) as measures of ‘quality of life’ (QL) (Prutkin & Feinstein, 2002). Therefore, instead of clear theoretical frameworks to guide questionnaire development, measures of QL are often based on a combination of measures that were developed for other purposes (Tennant, 1995).

The term ‘HRQL’ has arisen to conceptually clarify that the measurement of QL is restricted to those aspects of life that fall within the reach of the health care system. However, conceptual- or theory-based definitions of HRQL are rare (e.g., Wilson & Cleary, 1995). Moreover, the term HRQL is often used interchangeably with terms such as quality of life, wellbeing, perceived health, subjective health status, health perceptions, physical functioning, symptoms, or functioning disability. HRQL is broadly described as a multidimensional construct that constitutes physical, psychological and social wellbeing (based on the WHO definition of health). However, physical wellbeing includes concepts such as health status, functional capacity or physical symptoms. Psychological wellbeing may include depression, anxiety or cognitive functioning. And, social wellbeing can include family relations and social interaction. As a result, operationalizations of the HRQL construct can encompass one, but more often several of the constructs within the physical, psychological or social domains (for more detailed accounts of historical and theoretical developments of HRQL assessments see for example Bowling, 1997; Bullinger, 2002; Fayers & Hays, 2005; Fayers & Machin, 2016; Ferrans, 2005; Guyatt, Feeny & Patrick, 1993; McSweeney & Creer, 1995; Prutkin & Feinstein, 2002).

¹ Preamble to the Constitution of the World Health Organization as adopted by the International Health Conference, New York, 19-22 June, 1946; signed on 22 July 1946 by the representatives of 61 States (Official Records of the World Health Organization, no. 2, p. 100) and entered into effect on 7 April 1948.

Heterogeneity in operationalization of HRQL. As there is no conceptual consensus on what precisely constitutes HRQL, there is also no consensus on which components or domains should be used in its operationalization. For example, it has been argued that physical aspects of health are over-represented in HRQL assessments (e.g., Mayo, Moriello, Asano, van der Spuy, & Finch, 2011; Smith, Avis, & Assmann, 1999) and that there is no agreement about the mental or social aspects that should be considered (e.g., Hunt, 1997; Prutkin & Feinstein, 2002). With a lack of theory-based definitions of HRQL, researchers often use operational definitions of HRQL instead, i.e. they define the concept through the measurement instruments that are applied.

Barofsky (2012) argues that operational definitions may be required to understand an inherently abstract concept such as HRQL. That is, the construct has to be made concrete for it to make sense. Such practice may do justice to the different definitions of HRQL that are adopted for different groups of patients with different diseases, or with different treatments. As such, operational definitions of HRQL have led to the development of disease-specific HRQL questionnaires. These questionnaires focus on problems associated with specific disease status, patient groups, or areas of functioning. Generic HRQL instruments, on the other hand, are devised to provide a general summary of HRQL that can be used in a variety of contexts. For example, the EQ-5D (The EuroQol Group, 1990) and SF-36 health survey (Ware, Snow, Kosinski, & Gandek, 1993) were developed to be applicable to a wide range of health conditions and treatments, whereas the EORTC QOL Group has developed a generic quality-of-life instrument for cancer patients (the QLQ-C30) that can be complemented with tumor-specific modules (Aaronson, et al., 1993). Such developments have helped to provide some structure in the conceptualization – through operationalization – of the HRQL construct.

The heterogeneity in the measurements of HRQL does not only stem from differences in its (operational) conceptualizations, but also from differences in how exactly patients' HRQL is assessed. That is, two instruments may provide conceptually similar definitions of HRQL (e.g., both capture physical and psychological symptoms), yet they may differ in terms of how they operationalize (measure) these concepts with HRQL-items. These differences are mostly due to diverging interpretations of the term 'subjective' in 'subjective measurement'. When one considers the distinction between objective versus subjective measurements of health, objective measurements refer to clinical diagnostics², while subjective measurements refer to health as experienced by the patient (i.e., HRQL). Two patients may have identical objective health (i.e., identical clinical diagnosis or conditions) but can have very different subjective health experiences. The term subjective thus indicates that the meaning of health to the patient is taken into account. However, the term subjective is sometimes taken to only indicate that health is patient-reported, i.e. conditions are measured by asking the patient. Yet, as Mor &

² Note that one may question the objectivity of clinical assessments of health. Clinical assessments often require human interpretation in order to acquire meaning. Just because the observation is made by someone other than the patient, does not necessarily make the observation objective. See Cella et al. (2003) for a discussion on the accuracy and precision of HRQL data as compared to clinical measures.

Guadagnoli (1988) argue, a symptom or feeling that is experienced by the patient can still be seen as objective when it refers to a state or condition that does not require an evaluative process to determine the meaning of that symptom or feeling to that person (e.g., are you able to walk a 100 meters?). In contrast, when the patient is being asked to make a relativistic assessment (e.g., compared to three weeks ago, ...) or assess the meaning of the symptom (e.g., how much did the pain trouble you?), the evaluation may be considered truly subjective because it incorporates a personal anchor or frame of reference (see also Schwartz & Rapkin, 2004).

The distinction between objective and subjective patient-evaluations of health is important as it influences the interpretation of HRQL (change) assessments. Specifically, different types of health evaluation may be susceptible to different types of change. An objective evaluation such as ‘are you able to walk a 100 meters?’, might be less susceptible to response shift as the patient’s frame of reference is not expected to change (much) across occasions (i.e., the interpretation of what exactly is a 100 meters is expected to stay more or less the same). However, if we consider questions such as ‘are you able to make a *short* walk?’, or ‘are you *limited* in your ability to walk?’, the frame of reference of the patient may play a role in answering the question (i.e., terms such as ‘short’ and ‘limited’ leave room for personal interpretation). For example, Taminiu-Bloem et al. (2010) conducted cognitive interviews with a group of cancer patients when they filled out a HRQL-questionnaire, and found that a ‘short walk’ is interpreted to refer to a longer walk before radiotherapy than afterwards. More generally, the operationalization of health aspects as ‘health problems’ (i.e., objective evaluations) or ‘health evaluations’ (i.e., subjective evaluations) have been found to be differently susceptible to change (e.g., de Haes, de Duijter, Tempelaar, & Pennink, 1992; Hyland, Kenyon, & Jacobs, 1994); which could (partly) be due to response shift. Therefore, the heterogeneity in operationalization of the HRQL construct – and the distinction between objective and subjective health evaluations – is important to consider as it may influence the occurrence of different types of change.

Conclusion. HRQL outcomes provide valuable information about the patient-perspective on health that can complement and broaden standard clinical outcomes. However, it is important to consider the heterogeneity in conceptualization and operationalization of the HRQL construct. This will enable appropriate interpretation of HRQL outcomes, as it may influence the occurrence and interpretation of different types of change. As adaptation of a personal frame of reference is considered to be a key characteristic of HRQL assessment, the investigation of change in these frame of references, i.e. response shift, is especially relevant for a valid assessment of change in HRQL.

Typology of Change: Using Structural Equation Modeling to Investigate Response Shift

Structural equation modeling (SEM) is a statistical technique that can be used to model relationships between observed responses (e.g., patients’ scores on items or subscales of a HRQL questionnaire) to be reflective of one or more unobserved latent variables or common factors (e.g., the HRQL-construct that the items or subscales aim to measure). Within the SEM

framework, the variances and covariances (Σ , ‘Sigma’) and means (μ , ‘mu’) of the observed variables (X) are given by:

$$\text{Cov}(X, X') = \Sigma = \Lambda \Phi \Lambda' + \Theta,$$

and:

$$\text{Mean}(X) = \mu = \tau + \Lambda \kappa,$$

where Λ is a matrix of common factor loadings that describes the relationships between the observed variables and underlying common factors, Φ is a matrix of common factor variances and covariances, Θ is a matrix of residual variances and covariances that cannot be explained by the underlying common factors, τ is a vector of intercept values of the observed variables, and κ is a vector of common factor means.

Assessment of change. SEM can be applied to data from multiple measurement occasions, which enables the assessment of change. SEM provides a convenient framework for the assessment of change, as all model parameters can reflect (possible) change. Hence, it can be used to operationalize different types of change (Oort, 2004). The measurement issue of response shift can be approached with the SEM framework by using the concept of measurement bias (Oort, 2005). Unbiased measurement, or measurement invariance, entails that the observed variables measure the same latent variable in the same way across different situations (Meredith & Teresi, 2006). A violation of measurement invariance indicates that the comparison of underlying latent variables is biased. A formal definition of measurement invariance was first given by Mellenbergh (1989; but see also Meredith, 1993; Meredith & Millsap, 1992), where an observed variable X (e.g., an item or subscale from a HRQL questionnaire) measuring trait T (e.g., HRQL) is unbiased with respect to another variable V (e.g., disease, treatment, gender), if and only if:

$$f_1(X | V = v, T = t) = f_2(X | T = t),$$

where f_1 is the distribution of the observed responses given the values v and t of variables V and T , and f_2 is the distribution of observed responses given only the values t of variable T . When these distributions are not equivalent, the responses of X are biased with respect to V . In the presence of bias, differences between two people on observed scores may not reflect ‘true’ differences on the trait variable (e.g., men and women may score differently on an item that measures wellbeing, even though their wellbeing does not differ).

Although the concept of measurement invariance has been mostly applied to investigate across group differences (i.e., with V being group membership), Mellenbergh emphasized the generality of the definition, where V can be latent or manifest, and may have nominal, ordinal or interval measurement scales. When one replaces V with the time of measurement, it can be used to investigate longitudinal measurement invariance. In the common factor model, longitudinal measurement invariance holds when the measurement parameters in τ , Λ , and Θ are invariant

across measurement occasions. We can distinguish between different levels of measurement invariance, where *configural invariance* refers to invariance of the pattern of common factor loadings (i.e., the pattern of zero and non-zero values in Λ), *weak factorial invariance* refers to the invariance of both the pattern of common factor loadings and the values of the common factor loadings (i.e., all elements of Λ), *strong factorial invariance* refers to the invariance of common factor loadings and intercepts (i.e., Λ and τ), and *strict factorial invariance* refers to the invariance of common factor loadings, intercepts, and residual variances (i.e., Λ , τ , and the diagonal of Θ).

Operationalization response shift effects. The SEM method for the investigation of different types of changes in HRQL outcomes uses the notions of configural invariance, weak factorial invariance, strong factorial invariance, and strict factorial invariance to operationalize reconceptualization, reprioritization, uniform recalibration and non-uniform recalibration respectively. To illustrate, suppose we have patients' scores on several observed indicators that measure physical, emotional and social aspects of health. We use a (three-) factor model to represent the relationships between the observed variables, where the underlying latent variables represent everything that these measures have in common (e.g., perceived health or HRQL). The pattern of common factor loadings is used to determine whether patients employ the same conceptual framework across time. For example, when the common factor loading of an observed indicator of physical health would be zero at the second measurement occasion, this means that this indicator is no longer part of the measurement of physical health, indicating a shift in the conceptualization of physical health (i.e., reconceptualization response shift). When the value of the common factor loading is smaller at the second measurement occasion as compared to the first measurement occasion, this indicates a shift in the meaning (importance) of the indicator to the measurement of physical health (i.e., reprioritization response shift). When the intercept value changes over time, this indicates a shift in the meaning of the response categories (internal standards) of the indicator by which patients assess their physical health (i.e., uniform recalibration response shift). When the variances of the residual factors change over time, this may indicate that the change in the meaning of the response categories of an observed indicator is not in the same direction for all categories, or the direction of change differs between individuals (i.e., non-uniform recalibration response shift).

Operationalization 'true' change. In the presence of reconceptualization, reprioritization, or recalibration responses shift, the comparison between underlying latent factors (i.e., physical, emotional and social domains of HRQL) is compromised because the meaning of the construct is not equivalent across time. Moreover, a comparison of the indicators for which response shift has been detected is also compromised, as change in the observed indicators does not (only) reflect change in the underlying variables. The SEM framework for the investigation of change in HRQL outcomes does not only allow for the detection of possible response shifts, but also for the incorporation of detected response shifts into the model. By allowing some non-invariance of model parameters (i.e., partial measurement invariance; Byrne, Shavelson, & Muthén, 1989)

it is possible to investigate change in the underlying latent variables, while taking into account possible response shifts. Changes in the common factor variances (i.e., the diagonal of Φ) and common factor means (i.e., κ) across occasions are indicative of 'true' change in the construct of interest (i.e., the underlying latent variables that represent HRQL). When the common factor means increase over time, this indicates that patients' score higher on HRQL after the start of treatment.³ When the common factor variances increase, this indicates that the group of patients becomes more heterogeneous with regards to HRQL.

Added value of the SEM approach. There are two main advantages of the SEM approach to investigate change in HRQL outcomes. First, it allows for an operationalization of different types of response shift. Second, it can account for the different types of response shift by modeling partial measurement invariance. Investigation of change using the SEM framework thus allows for an investigation of change in HRQL where patients' self-evaluations are modelled to come from the same frame of reference, while deviations from this frame of reference (i.e., changes in standards, values or conceptualizations) are also considered. In addition, the flexibility of the SEM framework enables the inclusion of multiple measurements (e.g., analyses of more extensive follow-up trials), multiple groups (e.g., the comparison of different patient-groups based on disease, treatment, or patient-characteristics), or multidimensional scales (e.g., analyses including multiple HRQL domains, or other latent variables, simultaneously). Thus, the SEM method provides an elegant way for the investigation of different types of change in patient-reported health outcomes. However, like any method, its validity depends on certain methodological and conceptual assumptions. Detailed overviews of the general assumptions of SEM are provided elsewhere (e.g., Bentler & Chou, 1987; Bollen, 1989; Bullock, Harlow, & Mulaik, 1994) and include some more critical notes (e.g., Borsboom, Mellenberg, & van Heerden, 2003; Cohen, Cohen, Teresi, Marchi, & Velez, 1990). Below, we briefly discuss some of these issues, particularly those that are deemed important for the interpretation of change in HRQL outcomes.

HRQL as a latent variable. An underlying assumption of SEM is that the relationships between observed variables can be attributed to an underlying latent variable that is not directly observed (Bentler, 1982). In the case of HRQL, we thus assume that the construct itself cannot be directly observed, but instead we have observations that are *reflective* of HRQL. Reflective measurement assumes a directionality of effects between HRQL and the observed variables that has been much debated (e.g., Costa, 2015; Fayers & Hand, 1997; Fayers, Hand, Bjordal, & Groenvold, 1997). With reflective measurement, it is assumed that variations in the underlying latent variable are the cause of variations on the observed variables (Bollen, 1989). For example, patients score differently on questions about the emotional aspects of health *because* they have

³ As residual variances do not feature in the mean structure of the SEM model, the detection of non-uniform recalibration is not important for the investigation of mean change in the common factor (Oort, 2005; Vandenberg & Lance, 2000; but for an alternative view see also Lubke & Dolan, 2003; Deshon, 2004).

different HRQL. However, it could be argued that the direction of effects is reversed, where patients have a different HRQL because they have different emotional health, also referred to as *formative* measurement. With formative measurement, HRQL is considered an emergent property from its measurements, whereas with reflective measurement HRQL is seen as something that exists in reality, independent from its measurements. When applying the SEM framework to HRQL data, we should therefore consider whether the adopted HRQL construct is consistent with reflective measurement.

With a lack of theoretical frameworks of HRQL it is difficult to find theoretical or conceptual arguments in the literature that support the notion of the HRQL construct as a latent variable. In the literature on subjective wellbeing, causal relations between the concept and its measures have been discussed in terms of bottom-up versus top-down theories (Diener, 1984). One of these conceptual frameworks describes wellbeing as a top-down process, where the wellbeing of an individual determines individual perceptions of different aspects of life. If we apply this example to the HRQL construct, and conceive the individual perceptions of different health-aspects of life as being determined by an individual's HRQL, then this would be in line with reflective measurement. Similarly, in a discussion on the difference between reflective and formative measurement, Cohen et al. (1990) give an example of reflective measurement of (ill) health. They considered indicators such as fatigue and pain to be functional consequences of diseases and thus in line with reflective measurement of health. These examples of reflective indicators of health are consistent with common measures of HRQL. If we consider HRQL as perceived health by the patient, then we could argue that measures of specific areas of health (e.g., perceived pain, perceived fatigue) are functional consequences of HRQL.

However, the appropriateness of reflective measurement depends also on the operationalization of the HRQL construct. For example, in the same discussion of reflective and formative measurement by Cohen et al. (1990), they consider indicators such as blood-pressure or heart-rate as formative indicators of health. They argue that these kind of 'objective' clinical conditions cannot be conceived as being determined by health, but rather as causes of (ill) health. This argument has been raised too, specifically for the measurement of HRQL (Fayers & Hand, 1997; Fayers et al., 1997). They argue that common 'objective' indicators such as symptoms or side-effects cannot be considered reflective indicators of HRQL, but instead are 'causes' of HRQL. The theoretical discussion on subjective versus objective indicators of HRQL is thus – again – important to consider. One may feel comfortable with reflective measurement of HRQL, when it is measured through subjective evaluations of health aspects. In contrast, when HRQL is measured through objective indicators such as ability to walk, it is harder to conceive these indicators as consistent with reflective measurement. HRQL does not *determine* patients' ability to walk. Rather, patients' ability to walk determines their HRQL. However, one should take into account to what extent these 'objective' indicators are asking for an evaluation from the patient. For example, when symptoms or side-effects require an evaluation in terms of severity or impact, patients' answers may be influenced by their health perception or HRQL.

One can argue that patients' answers to questions about health will at least partly reflect their underlying HRQL (Oort, 2005).

The lack of consensus on the issue of formative versus reflective measurement is also related to the lack of consensus on the definition of the HRQL concept. One may – perhaps rightfully – question whether HRQL can be considered an existing entity on itself, rather than being a phenomenon that exists only because of pragmatic and functional utility in explaining correlational patterns. Determining the general appropriateness of reflective measurement for the HRQL construct is obfuscated by its different conceptualizations and operationalizations (Costa, 2015). The view in this thesis is that SEM provides a convenient framework to model commonalities in the patterns of change of multiple health-indicators, but that the (successful) application of these models should not be interpreted as evidence that the reflective latent-variable model is an appropriate or 'true' model for HRQL. This may be an inconsistent stance, as the SEM model is formulated under the assumption that the model is 'true', and therefore the model (parameters) may only be valid under this assumption. Nevertheless, the possible methodological and practical benefits of investigating change in HRQL outcomes using SEM methodology are deemed important to further develop the field of HRQL research. Theoretical developments are imperative to reach agreement on the conceptualization and operationalization of HRQL. Statistical methods may help to understand the behavior of HRQL measurements – in its current stage of the theoretical development – and therefore aid also in its conceptualization.

Interpretation of measurement bias. The different types of change in HRQL outcomes are operationalized using measurement invariance and measurement bias. The term 'bias' has a negative connotation, and is often referred to as 'confounding', requiring 'adjustment', or 'obfuscating true change'. This has led to some disagreement in the scientific community about whether or not response shift effects should be interpreted as 'true' changes in HRQL, or whether they should be interpreted as a 'bias' in the measurement of HRQL (Kievit, et al., 2010; Ubel, Peeters, & Smith, 2010). Some disagreement in the interpretation of response shift stems from the fact that response shift can be viewed from a measurement perspective, but also from a conceptual perspective, where response shift is considered a bias in the explanation of change (Oort, Visser, & Sprangers, 2009). It is important to understand that the term bias is used to refer to measurement bias, specifically in the comparison of HRQL assessments. It should not be taken to indicate that the HRQL assessment of one specific occasion is biased, but rather that the comparison of HRQL assessments is biased when one cannot ascertain that the same construct is measured at both occasions. Thus, the term bias only refers to bias in terms of the measurement of change. Although bias may confound or obfuscate assessment of change from a measurement perspective, the occurrence of detected bias may also enrich our understanding of patterns of change. In fact, it has been emphasized that the interpretation and explanation of detected measurement bias is of utmost importance – although somewhat neglected in the literature (e.g., Mellenbergh, 1989; Millsap & Meredith, 2004). The SEM framework for the

investigation of change in HRQL outcomes is innovative in that it formulates a direct link between statistical operationalizations of measurement bias and the conceptualizations of change due to response shift effects. Response shift effects are often acknowledged to be of substantive interest (e.g., Barclay-Goddard, Epstein, & Mayo, 2009) and may even be considered as beneficial treatment effects (Nolte, Elsworth, Sinclair, & Osborn, 2012; Preston et al., 2013). Therefore, insight into the changes in standards, values and conceptualizations of patients' HRQL may also enhance our understanding of changes in HRQL.

Group-level inferences. SEM is applied to group-level statistics (i.e., variances, covariances and means of the observed indicators) and provides information at the group level (i.e., variances, covariances and means of the common factors). There is some discussion as to whether the concept of HRQL should be seen as something that is only inherently meaningful to an individual patient, or whether HRQL scores can also be meaningfully aggregated to represent an 'average' of groups of patients (e.g., Donaldson, 2005; Hunt, 1997; Lantos, 1998). In the assessment of change, these average scores of HRQL assessments are interpreted relatively to the average scores of another HRQL assessment in order to provide information about group-level change. It is true that group-level change may not be directly meaningful for inference about individual-level change (e.g., some patients may show no change or even negative change). The same is true for detecting changes due to response-shift effects. The SEM method will only detect group-level response shift when the majority of patients show individual-level response shift (Oort, 2005). When the majority of patients experience a response shift, this can be meaningful for the interpretation of general patterns of change – even though 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 these findings may show which patients are prone to show the detected response shift, may help to understand why certain patients do or do not experience the response shift, and possibly how response shift could be enhanced or prevented.

The SEM method for the investigation of response shift is thus valid under the assumption that response shift is present in the majority of individuals. However, the interpretation of group-level patterns of change as indicative of individual-level patterns of change in the majority of patients requires the assumption of 'ergodicity' (Molenaar, 2004). Ergodicity implies that intraindividual differences (i.e., changes at the individual level) have the same structure as interindividual differences (i.e., changes at the group level). It has been argued that ergodicity is an unrealistic assumption for the majority of psychological constructs (Borsboom, et al., 2003), and specifically for the construct of HRQL (Donaldson, 2005). For example, it may be that individual patients have different latent variable structures of HRQL (e.g., with different dimensional structures), which may yield different patterns of change. Such individual differences may not be adequately captured when analyzing the latent variable structures of groups of patients. Empirical evidence for similarities between inter- and intra-individual structures of HRQL does not yet exist. Some promising research has been done to identify cognitive processes at the individual level that may be used to substantiate the interindividual

conceptualization of HRQL (change) assessments (Bloem, 2010). The individual cognitive processes used to elicit responses to HRQL-questionnaire items were found to change over time, and could be linked to processes of response shift (Taminiau-Bloem, et al., 2010). However, Taminiau-Bloem et al. also found substantial intra- and interindividual differences in the patterns of cognitive processes. Further research is needed to connect cognitive processes of the individual with group-level structures of (change in) HRQL. Moreover, statistical analyses of processes of change and structures of HRQL in individuals can be used to investigate the equivalence with processes of change and structures of HRQL in groups of patients, and thus test the assumption of ergodicity.

The SEM method is applied under the assumption that the latent variable structure of HRQL at the group-level also applies ‘on average’ to the individual-level or to the majority of individuals. This is necessary for a meaningful interpretation of group-level change. However, one should keep in mind that SEM is not directly aimed at making inferences about processes of change (or response shift) at the level of unique individuals. Alternative statistical techniques are available to study mechanisms of change within individuals directly (e.g., Hamaker, Dolan, & Molenaar, 2005).

Conclusion. The SEM framework can be used to assess different types of change in HRQL outcomes. Specifically, the concept of measurement bias can be used as an operationalization for response shift effects, which enables the distinction between changes in standards, values, and conceptualization of patients’ self-evaluation, and changes in the underlying construct of interest (e.g., HRQL). However, the heterogeneity in conceptualization and operationalization of the concept of HRQL complicates the interpretation of change. The tenability of the underlying assumption that HRQL is a latent variable that refers to an existing, but unobserved, entity depends (partly) on the definition and operationalization of HRQL. In this thesis, the SEM method to assess change in HRQL outcomes is not used in order to claim conceptual evidence of HRQL as a latent variable, but rather to investigate patterns of change in HRQL measurements and deviations from these patterns of change that can be taken as evidence of response shifts. Even though an individual’s quality of life is an idiosyncratic response of which the meaning to that individual’s life may only be understood using qualitative or phenomenological approaches, we can use these individual patient responses in a quantitative way to uncover commonalities in the patterns of change of perceived health. This may give valuable insight into the patterns of experienced illness and disease of groups of patients, which may help our understanding of changes in patient-reported quality of life, and eventually may also benefit the quality of life of the individual patient.

Thesis Outline

The general aim of this thesis is to help researchers and practitioners with the investigation and interpretation of change and response shift effects in HRQL outcomes. The following chapters address several methodological issues that are important for the assessment of different types of change in HRQL outcomes using the proposed SEM approach. First, we compare the SEM

approach to the ‘then-test’ approach, which is one of the most commonly applied methods for the detection of response shift (Chapter 2). Then, we explain how the SEM approach for detection of response shift can be extended to the situation in which there are many measurement occasions (Chapters 3, 4 and 5), and to the analysis of discrete data (Chapters 6 and 7). Finally, we explain how to calculate and interpret effect-size indices of change to enable interpretation of the clinical significance of response shift (Chapter 8).

Comparison with the then-test approach. There are different methodological approaches available for the investigation of change due to response shifts. A commonly applied method for the detection of response shift is the so-called ‘then-test’ approach. This design approach extends the pretest-posttest design with a retrospective pre-test at time of post-test assessment, where patients are being asked to re-evaluate their HRQL at time of pre-test. In comparison to the SEM approach, the then-test approach thus requires an additional measurement and only measures recalibration response shift. In [Chapter 2](#) we compare both approaches in terms of their assessment of change and detection of response shift. In addition, the inclusion of the then-test into the SEM approach enables the evaluation of underlying assumptions of the then-test. The aim of this chapter is to test the validity of the then-test approach and highlight the differences between the then-test and SEM approach for the investigation of change. Both approaches are applied to data from 170 cancer patients undergoing invasive surgery. HRQL was assessed prior to surgery (pre-test) and three months following surgery (post-test and then-test) using the SF-36 health survey (Ware et al., 1993) and the Multidimensional Fatigue Inventory (Smets, Garssen, Bonke, & de Haes, 1995).

Investigation of response shift in extensive longitudinal designs. Response shift is usually investigated in situations where there are two measurement occasions (i.e., a pre- and post-test). The flexibility of the SEM framework allows for the inclusion of many more measurement occasions. However, this requires additional restrictions to the model. Imposition of so-called Kronecker product restrictions yield the longitudinal three-model model (L3MM; Oort, 2001) which enables the assessment of change in more extensive longitudinal designs. In [Chapter 3](#) we explain how the L3MM can be applied to multivariate longitudinal data from many measurement occasions, and illustrate the imposition and interpretation of L3MM restrictions. In [Chapter 4](#) we propose a procedure for the detection of measurement bias in L3MMs, and explain how the detected biases can be modelled using linear, or non-linear curves. This enables the detection of response shift in L3MMs and the evaluation and interpretation of possible trends in response shift effects over time. In both chapters we use an illustrative example of HRQL data that was obtained from 682 patients with painful bone metastasis at 13 measurement occasions; before and every week after treatment with radiotherapy. This is a subset of data from the Dutch Bone Metastasis Study (DBMS; Steenland et al., 1999; van der Linden et al., 2004), where HRQL was assessed with the EQ-5D (The EuroQol Group, 1990),

the Rotterdam Symptom Checklist (de Haes et al., 1996), and the QLQ-C30 (Aaronson et al., 1993). In [Chapter 5](#) we apply L3MMs in a multi-group context to investigate response shift and assess change in groups of patients with different attrition rates. We included data from 1029 patients from the DBMS database, and distinguished three groups based on their pattern of attrition: short survival (3-5 measurements; $n = 144$), medium survival (6-12 measurements; $n = 203$), and long survival (>12 measurements; $n = 682$).

Investigation of response shift in discrete data. SEM is usually applied to continuous data, i.e., modeling the means, variances and covariances of observed variables. However, HRQL-data are often not continuous but discrete (e.g., ordinal item responses). As a consequence, SEM is generally applied to continuous item responses or to the aggregated sum of ordinal item responses (i.e., at the subscale level). In [Chapter 6](#) we propose a SEM approach for the investigation of response shift and assessment of change in discrete data. The proposed SEM approach is illustrated with item-level data from 485 cancer patients whose HRQL was measured with the SF-36 health survey, before and after start of chemo- or radiotherapy. In [Chapter 7](#) the SEM approach for discrete data is applied in a multi-group context to investigate gender- and age-related bias in the items of the Hospital Anxiety and Depression Scale (HADS; Zigmond & Snaith, 1983). Data were obtained from 1068 adults who consulted a primary care professional. We illustrate bias detection using a multigroup SEM approach and a multidimensional SEM approach, and compare our results to the results of the ordinal logistic regression, item response theory, and contingency tables methods reported by Cameron, Scott, Adler and Reid (2014).

The clinical significance of response shift. The detection of response shift is guided by tests of statistical significance. However, when an effect is statistically significant this cannot be taken to indicate that the effect is also clinically significant (i.e., meaningful). In [Chapter 8](#) we explain how SEM can be used for the decomposition of change, where observed change is decomposed into change due to response shift effects, and change due to the underlying latent variable (e.g., HRQL). Subsequently, we explain how to calculate and interpret effect-size indices of change to enable interpretation of the clinical significance of the different types of change. As such, effect-size indices can be used to evaluate the impact of response shift on the assessment of change. To further enhance clinical interpretability, we compare the effect-size indices of change with other well-known types of effect-size indices, including probability benefit, probability net benefit, and number needed to treat to benefit. Pre- and post-test data from [Chapter 2](#) are used as an illustrative example.

Summary. In [Chapter 9](#), we provide a summary of the main findings of this thesis. In addition, we discuss practical issues that play a role in the application of the SEM approach for the detection of response shift, and provide guidelines for its future applications in the context of HRQL research.