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

Verdam, M.G.E.

# CHAPTER 9

## Summary and General Discussion

Structural equation modeling (SEM) can be used to investigate different types of change in health-related quality of life (HRQL) outcomes, among which so-called 'response shift' (Oort, 2005). Response shift refers to a change in the meaning of one's self-evaluation, caused by a change in internal standards of measurement (i.e., recalibration), values regarding the relative importance of subdomains (i.e., reprioritization), or definition of the target construct (i.e., reconceptualization; Sprangers & Schwartz, 1999). Response shift effects may cause changes in observed scores that are not directly caused by changes in the construct of interest (e.g., HRQL). Therefore, taking into account possible response shift is important for a valid assessment of change. The overall aim of this thesis is to facilitate applications of response shift detection, and thereby contribute to a better understanding of response shift phenomena and thus change in HRQL. In the next section, we summarize the main findings of this thesis, 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.

**Summary**

This thesis focused on several methodological issues with regard to the SEM approach for the detection of response shift and the assessment of change in HRQL outcomes. We compared 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). We extended the SEM approach for detection of response shift to the situation in which there are many measurement occasions (Chapters 3, 4 and 5), and for the analysis of discrete data (Chapters 6 and 7). Finally, we explained how to calculate and interpret effect-size indices of change to enable interpretation of the clinical significance of response shift (Chapter 8). Below, we highlight the main contributions of this thesis by summarizing the features of the SEM approach that were developed, investigated, and illustrated.

**Comparison with the then-test approach.** In Chapter 2 we compared the SEM approach for the detection of response shift to the so-called 'then test' approach. The then-test approach includes 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. The aim of this chapter was to compare both methods in terms of their approach to response shift detection and assessment of change. In addition, inclusion of the then-test into the SEM approach was used to evaluate the underlying assumptions of the then-test. The then-test approach is designed to detect only recalibration response shift (i.e., assuming that all response shift is of the recalibration type; recalibration assumption), it measures 'true' change by comparing post-test and then-test assessments (i.e., assuming that they are completed with the same frame of reference; consistency assumption), and assumes that patients are able to correctly recall their state at pre-test (i.e., recall assumption). Compared to the then-test approach, advantages of the SEM approach include the possibility to detect response shift without the need of additional measurements, and the possibility to differentiate between different types of response shift. Both approaches were applied to HRQL-data from 170 cancer patients undergoing invasive surgery, where HRQL was assessed prior to surgery (pre-test) and three months following surgery (post-test and then-test) using the SF-36 Health Survey and the Multidimensional Fatigue Inventory. The SEM approach identified the occurrence of not only recalibration, but also reprioritization and reconceptualization response shift (i.e., violating the recalibration assumption). In addition, we found that the frames of reference were not invariant across post- and then-test assessments (i.e., violating the consistency assumption). However, we did not identify recall bias (i.e., supporting the recall assumption). Furthermore, we found that both approaches revealed a similar pattern of change, but that the SEM approach detected more response shift effects which resulted in some differences in the size and direction of change. In this chapter we showed that SEM can be used to make a substantive comparison between different methodologies for the detection of response shift. In addition, testing the underlying assumptions of the then-test approach through SEM is useful for determining the validity of the then-test approach.

**Investigation of response shift in extensive longitudinal designs.** Response shift is usually investigated in the situation where there are only two measurements, i.e., a pre- and post-test. However, longitudinal clinical trials often include many more measurement occasions (e.g., extensive follow-up measures to evaluate long-term effects). To facilitate the analysis of data from such longitudinal designs, we extended the SEM approach for the investigation of change in multivariate longitudinal data by adaptation of the 'Longitudinal Three-Mode Model' (L3MM). In Chapter 3 we explained how to impose L3MM restrictions on the parameter matrices of the SEM model. The L3MM restrictions substantially reduce the number of parameter estimates (i.e., leading to more parsimonious models), and yield separate estimates for the relationships between variables and the change in the relationships between the variables over time. The assessment of change with L3MMs was illustrated using 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 was a subset of data from the Dutch Bone Metastasis Study (DBM), where HRQL was assessed with the EQ-5D, the Rotterdam Symptom Checklist, and the EORTC QLQ-C30. Results indicated that correlations between variables from the first and second measurement occasion were smaller than the correlations between variables of the second and third measurement occasion, and so on. This pattern of change suggests that patients became more homogenous in their answers to the questionnaires. In addition, we illustrated how to test substantive hypotheses about change in the correlational patterns between variables, and change in the means of the underlying factors (e.g., HRQL). Thus, in this chapter we showed that L3MMs do not only facilitate the analysis of complex longitudinal data but also the substantive interpretation of the dynamics of change.

In Chapter 4 we explained how to proceed with the investigation of measurement invariance in L3MMs. Specifically, additional parameter matrices were introduced to accommodate possible violations of measurement invariance and to enable the investigation of bias in individual factor loadings and intercepts. We applied the investigation of measurement invariance with L3MMs to the same HRQL-data that we used in Chapter 3, and showed that further investigation of cases of bias (i.e., response shift) is possible through modeling the measurement bias using linear and non-linear curves. We identified three cases of measurement bias and illustrated the interpretation of a linear, piece-wise linear, quadratic, and inverse trend. The proposed methods can thus be used to investigate trends in detected response shift effects, which will lead to more insight into the development of these effects. We concluded that L3MMs can advance the interpretation of findings from extensive longitudinal designs.

In Chapter 5, a multigroup SEM approach was used to investigate response shift effects in groups of patients with different patterns of missing data due to attrition. We used HRQL data from 1029 patients of the same DBMS database that was used in the previous two chapters, 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). Imposition of L3MM restrictions facilitated the analyses of HRQL-data from groups of patients who completed 3, 6 and 13 measurements respectively. In addition, the multigroup approach enabled the investigation of differences in the patterns of change between groups of patients. Results showed that patterns of change in HRQL differed among patients with short, medium or long survival. Moreover, the different groups of patients were not equally affected by the detected response shifts. Thus, in this chapter we showed that SEM can be used to detect response shift and asses change across a substantial number of measurements, but can also be used to investigate differences in response shift and change across groups of patients.

**Investigation of response shift in discrete data.** SEM is especially suited for the analysis of continuous data. However, HRQL-data are often not continuous but discrete. As a consequence, SEM is often applied to continuous item responses or to the aggregated sum of ordinal item responses (i.e., at the subscale level). To facilitate the application of SEM for response shift detection to other types of data, we extended the SEM approach to include a modeling stage in which the observed discrete ordinal variables are modelled to be reflective of underlying continuous variables (Stage 1). Stage 1 yields estimates of means and variances and covariances that can be used for the detection of response shift and assessment of true change in Stage 2. In Chapter 6 we explained the proposed SEM approach for the detection of response shift in discrete data, and illustrated the detection of response shift in the items of the SF-36. We used data from 485 cancer patients whose HRQL was measured before and after start of chemo- or radio-therapy. Response shift was detected in items from five out of eight subscales of the SF-36 questionnaire. Overall, patients' mental health improved, while their physical health, vitality, and social functioning deteriorated. No change was found for the other subscales of the SF-36. Thus, in this chapter we showed how the SEM approach for discrete data enables the investigation of response shift and assessment of change at the item-level. We concluded that the proposed approach may improve our understanding of the response shift phenomena and therefore enhance interpretation of change in the area of HRQL.

In Chapter 7 the SEM approach for discrete data was applied in a multigroup context to detect gender- and age-related bias in the Hospital Anxiety and Depression Scale. Data was obtained from 1068 patients who consulted a primary care professional. We showed two ways of using SEM to detect item bias: multigroup SEM, and multidimensional SEM. In addition, results were compared to the results of the ordinal logistic regression, item response theory, and contingency tables methods reported by Cameron, Scott, Adler and Reid (2014). Results from the SEM approaches generally agreed with the results of Cameron et al., although the SEM approaches identified more items as biased. In this chapter we showed that SEM provides a flexible tool for the investigation of item bias in health-related questionnaires. Advantages of the SEM approach are that it enables the detection of both uniform and nonuniform bias, and that possible item bias can be taken into account to assess true differences between groups. Moreover, multidimensional SEM has practical and statistical advantages over multigroup SEM, and over other item bias detection methods, as it enables item bias detection with respect to multiple variables, and of various measurement levels.

**The clinical significance of response shift.** Detection of response shift is usually guided by tests of statistical significance. However, statistical significance does not indicate that the detected effects are also clinically significant. Therefore, in Chapter 8 we explained how to calculate and interpret effect-size indices of change. Specifically, we used SEM for the decomposition of change, where observed change is decomposed into: 1) change due to recalibration response shift, 2) change due to reprioritization and/or reconceptualization response shift, and 3) change due to change in the construct of interest (e.g., HRQL; 'true' change). Subsequently, calculating effect-size indices of change (i.e., standardized response mean) enabled evaluation and interpretation of the size of response shift effects, and the impact of response shift on the assessment of change. To further enhance the clinical interpretability of effects we showed how the proposed effect size of change relates to other well-known types of effect-size indices, i.e., the probability benefit, the probability net benefit, and the number needed to treat to benefit. Pre- and post-test HRQL-data from Chapter 2 were used as an illustrative example for the calculation and interpretation of the proposed effect-size indices of change. The detected response shift had small effects on change in the observed indicators, and on overall change in HRQL. To conclude, in this chapter we showed how effect-size indices can be used to give an indication of the clinical significance of response shift, and their impact on the assessment of change in HRQL outcomes.

## Practical Issues in Application of the SEM Approach

In order to facilitate future applications of the SEM approach for the assessment of change and detection of response shift in HRQL outcomes, we discuss some practical issues that are important for application of the method and interpretation of results.

**Establishing an appropriate measurement model.** An appropriate measurement model is a prerequisite for the investigation of change and possible response shift effects. The measurement model includes the specification of relations between the observed variables and underlying latent factor(s), and thus defines the measurement structure of the data. With longitudinal data, the measurement model includes the specification of the measurement structure at each measurement occasion; sometimes referred to as the longitudinal measurement model. To arrive at the longitudinal measurement model one can establish an appropriate measurement model for each measurement occasion separately, and combine all separate measurement models into a single longitudinal measurement model. Or, alternatively, one can combine all measurement occasions into a single longitudinal measurement model, and establish an appropriate measurement model for all measurement occasions simultaneously. The only requirements of the specified (longitudinal) measurement model is that the measurement structure is largely the same across time, and that it has interpretable common factors. Ideally, the specification of the measurement model is based on theory about the measurement structure of the construct of interest. For example, HRQL is often described as a multidimensional construct that encompasses social, mental and physical aspects of health. When a HRQL questionnaire is developed based on this theoretical framework (i.e., the items reflect the three different domains of HRQL) then the measurement model could be specified as a three-factor model, where all items that share a domain load on the associated common factor. Specification of the measurement model can become more complicated in situations where the dimensional structure of a questionnaire is unclear, or where (items of) different questionnaires are combined. Moreover, it is often necessary to modify the initially specified measurement model to obtain a well-fitting model. A well-fitting measurement model is necessary, as the measurement model is the baseline model against which all further models (i.e., that include equality restrictions on model parameters) will be compared.

Establishing an appropriate measurement model often includes an exercise of exploratory nature. An appropriate starting point for the specification of a measurement model can be based on the structure of the questionnaire, results from previous research, substantive considerations about the content of the observed measures, exploratory factor analyses, or – more likely – a combination of these approaches. The measurement model represents the most parsimonious, substantively the most reasonable, and best fitting model to the data (Byrne, Shavelson, & Muthén, 1989). Statistical criteria can be used to evaluate whether the model fit of the measurement model is appropriate (e.g., overall model fit) and to guide model respecification when the initial model fit is not adequate (e.g., using differences in model fit). However, making a decision on which and how many model respecification are necessary requires substantive considerations (i.e., does a model make sense?). For example, statistical indices may indicate that the largest improvement in fit can be achieved by freeing a factor loading of a physical functioning item on a common factor that measures mental health; such a model respecification may not make sense substantively. On the other hand, freeing a residual covariance between indicators that share the same item format may be sensible even though it will not lead to large improvements in model fit. In order to find a substantively reasonable measurement model, it is at least equally – and possibly even more – important to rely on substantive knowledge and practical interpretation, as on statistical criteria. Thus, it is not only statistical criteria, but rather the combination of statistical and substantive criteria that should be used to substantiate – and judge – the appropriateness of the measurement model.

**Identification of possible response shift effects.** The overall occurrence of response shift is evaluated by comparing the model that includes equality restrictions on all model parameters associated with response shift to the measurement model; representing an 'omnibus test' for the presence of response shift. This procedure has also been advocated by others (Millsap, 2010), and has been shown to protect against false positives (Vanier, Sébille, Blanchin, Guilleux, & Hardouin, 2015). However, if there is evidence of the presence of response shift, how does one then accurately locate which observed variable is affected by which type of response shift?

The search for biased model parameters requires exploratory model-fitting or respecification, which is referred to as the 'specification search'. The specification search can be guided using statistical criteria, such as modification indices, expected parameter changes, Wald tests, or differences in model fit (Jöreskog & Sorbom, 1996). In order to correctly identify the biased model parameters, it has been recommended to use an iterative procedure (Cheung & Rensvold, 1999), where all model parameters associated with response shift are freed one at a time, and the freely estimated parameter that shows the largest improvement in model fit is incorporated in the model. However, it may be that two different model modifications lead to equivalent improvement in model fit. A decision on which model modification to prioritize can thus not be based on statistical criteria alone. Given the dependence of sequential model respecification, freeing one model parameter may render freeing the other model parameter unnecessary, i.e., a change to the model can affect other parts of the model too. It may therefore be possible that alternative series of model respecifications lead to different results.

The specification search for possible response shift effects also requires a decision on when to stop searching. The aim of the specification search is to identify all possible response shift effects. Meanwhile, however, one wants to prevent the identification of trivial differences in model parameters across time as being of substantive interest. In addition to the improvement in model fit for freeing individual parameters, one can rely on the difference in model fit between the measurement model and the model that includes all identified response shift effects. When the overall difference in fit between these models is not significant, this may be taken as an indication that freeing additional model parameters is no longer necessary. Also, one can use the overall model fit of the model to judge whether the model that includes response shift is tenable. These model fit evaluations may provide more robust stopping criteria. However, it has also been argued that in order to adequately identify all biased parameters it may be necessary to continue the specification search, even when the established model already shows adequate model fit (McCallum, 1986). Therefore, model fit criteria should be used in combination with substantive criteria with regard to the (possible) biased model parameters. For example, it may be that freeing an additional model parameter will lead to a small, significant improvement in model fit, but that the associated response shift does not have a clear interpretation. As a researcher, one has to find a balance between the goodness of fit and the interpretability of the model.

The data-driven exploratory nature of the specification search thus implies that the resulting models must be evaluated with caution (MacCallum, 1986). As long as one is aware of its exploratory character, the process can be meaningful (Byrne et al., 1989) and will help to correctly identify response shift effects.

**Evaluation of overall model fit and difference in model fit.** The fit statistic to evaluate overall-goodness of fit is the chi-square test of 'exact' fit, where a significant chi-square value indicates a significant deviation between the model and data. In addition, the chi-square values of two nested models (i.e., where the second model can be derived from the first model by imposing restrictions on model parameters) can be compared to test the difference in 'exact' fit. A significant difference in chi-square values indicates that the less restricted model fits the observed data significantly better than the more restricted model, or in other words, the more restricted model leads to a significant deterioration in model fit. The chi-square (difference) test thus has clear interpretation and provides a convenient decision rule for the evaluation of overall model fit (e.g., the appropriateness of the measurement model) and the difference in model fit between two nested models (e.g., the overall presence of response shift or the identification of specific response shift effects). However, the chi-square test of exact fit is highly dependent on sample size, i.e., with increasing sample size and equal degrees of freedom the chi-square value increases. In addition, it tends to favor highly parameterized models as the chi-square value decreases sharply when parameters are added to the model. Therefore, the *practical* usefulness of the chi-square (difference) test as a single decision rule to evaluate (differences in) model fit has been questioned (e.g., Wu, Li, & Zumbo, 2007).

As an alternative to the chi-square test of 'exact' fit, a variety of other fit indices have been developed that provide descriptive evaluations of model fit. Examples include the root mean square error of approximation (RMSEA; Steiger & Lind, 1980; Steiger, 1990), the comparative fit index (CFI; Bentler, 1990), and the expected cross-validation index (ECVI; Browne & Cudeck, 1989). These indices of so-called approximate fit are less dependent on sample size and reward model parsimony. However, as the sampling distributions of many of these approximate fit indices are unknown, they cannot be used for formal hypothesis testing (Wu, et al., 2007). Instead, numerous cut-off criteria have been proposed that serve as 'rules of thumb' for the evaluation of goodness of fit (e.g., Schermelleh-Engel, Moosbrugger, & Müller, 2003; Hu & Bentler, 1999; Marsh, Hau, & Wen, 2004). For example, RMSEA values below .05 are generally taken to indicate 'close' approximate fit, and values between .05 and .08 indicate 'reasonable' approximate fit (Browne & Cudeck, 1992). Similarly, cut-off criteria have been proposed for the evaluation of differences in model fit using the difference in approximate fit indices (Cheung

& Rensvold, 2002). However, it has been argued that these approximate fit indices can only provide a preliminary interpretation of (differences in) model fit, but that these 'rules of thumb' should not be used as decision rules (Marsch, et al. 2004). Two approximate fit indices can be regarded as an exception in this regard. The RMSEA and ECVI have known distributions and associated confidence intervals, that can be used to give an indication of the precision of the fit estimates. The RMSEA is especially informative for the evaluation of overall model fit, and can be used to provide a 'test of close fit' (Browne & Cudeck, 1992). The difference in ECVI values of two nested models may be used to test the equivalence in approximate model fit. However, stringent evaluation of the performance of the ECVI for the comparison of nested models has not yet been performed. Thus, application of the ECVI in evaluation of differences in model fit should be proceeded with caution.

Approximate fit indices can be used to provide (descriptive) evaluations of model fit. However, they are not a panacea for finding 'golden rules' against which to evaluate model fit or differences in model fit (Wu et al., 2007). Moreover, with different tests and indices available to evaluate model fit, providing decision rules on whether the fit of a model is 'good' is further complicated by the fact that one might find inconsistent results (e.g., a significant exact chi-square test, but close approximate fit according to the RMSEA). The evaluation of differences in model fit is complicated for the same reasons. For example, Wu et al. (2007) found that the evaluation of differences in model fit using differences in CFI values and chi-square values almost always lead to contradictory conclusions. The researcher thus has to make a decision about the fit index that is most appropriate for the data and hypotheses under study. For example, the chi-square test of exact fit may detect small, but trivial, differences between model and data when the sample size is large or when the model is parsimonious. On the other hand, in the specification search to identify response shift effects, it might be desirable to have a high power to detect differences in parameter estimates. In addition, approximate fit indices can provide an appropriate alternative to the test of exact fit in certain situations (e.g., when one wants to accept some misfit in favor of model parsimony), but the interpretation of approximate fit indices according to their 'rules of thumb' is not generalizable to all situations. For example, it has been shown that many of the approximate fit indices are still sensitive to model parsimony (Cheung & Rensvold, 2002; Marsh et. al., 2004). Therefore, one may want to apply more stringent cut-off criteria for parsimonious models, and relatively less stringent cut-off criteria for complex models (Wu, et al. 2007).

Not only sample size and model complexity may affect the appropriateness of the chosen model fit index, but the type of data that is being analyzed may also play a role. For example, for the analysis of discrete data one has to choose between alternative estimation methods (e.g., Muthén, 1983; Browne, 1984; Satorra & Bentler, 1988), which produce different versions of the chi-square test statistic. In addition, these different chi-square statistics may require adjustment for an appropriate evaluation of overall model fit or differences in model fit (e.g., Satorra & Bentler, 2001; Asparouhov & Muthén, 2006). As many of the approximate fit indices are derived from the chi-square value, the choice of the chi-square test statistic (adjustment) may thus also influence the calculation and subsequent interpretation of the descriptive fit indices.

In general, decisions on model fit and differences in model fit require a careful consideration of the data, the model, the sample size, and the hypothesis that is being tested. Therefore, substantive decisions play an important role in the evaluation of (differences in) model fit, and might even require the use of different fit indices or different decision rules that are compatible with the purpose of the analysis.

**Interpretation and explanation of detected response shift.** With SEM, the concepts of measurement invariance and measurement bias are used to operationalize response shift effects. That is, when a measurement parameter shows measurement bias across time, this is taken as evidence of response shift. By using the non-invariance of measurement parameters to identify response shift effects, we do not look at response shifts directly, but at the effects these response shifts have on the measurement of HRQL. This allows us to describe *what* occurs (i.e., patterns of change), but it does not imply that we also know *how* it occurs (i.e., the cause of the identified bias). For the substantive interpretation of change it is therefore important to provide an interpretation and possible explanation of detected response shift. For example, imagine that recalibration was detected in the indicator 'pain' of perceived physical health, where patients showed a larger decrease in pain as compared to the other indicators of physical health. A possible explanation for this result may be that patients adapted to the experience of pain and therefore rated their pain to be lower at post-test, even though their actual experience of pain did not change (or changed to a lesser degree). It may also be that patients received treatment or medication that reduced their experienced level of pain. When the decrease in experienced pain was larger than the decrease on the other indicators of physical health, this effect may also be identified as recalibration response shift. However, one could argue that only the first interpretation coincides with what Sprangers & Schwartz (1999) describe as recalibration response shift. It is true that the SEM approach for the detection of response shift allows for a broad definition of response shift, as it includes all potential explanations for the changes that occur between measurements. Therefore, substantive interpretation of detected response shift is of paramount importance; it is needed to exclude, or make less likely, alternative explanations.

The interpretation of detected response shift can be based on substantive knowledge of the patient group, the treatment or disease trajectory. In addition, it is possible to include operationalizations of potential predictors of response shift in the SEM model. If measures of antecedents (e.g., sociodemographic or personality characteristics) or mechanisms (e.g., coping strategies, social comparison) are available, they can be incorporated in the model as possible explanatory variables for the detected response shift effects. Alternatively, qualitative research can be used to identify the cognitive processes that play a role in patients' self-evaluations, which in turn may provide valuable information on the occurrence and explanations for response shift (e.g., Taminiau-Bloem et al., 2016). Substantive interpretation and explanation of response shift is necessary to connect the detected response shift with experiences of response shift by patients. As such, it may help to substantiate the meaningfulness of response shift, and change in HRQL.

**Recommendations.** For researchers and practitioners who will apply the SEM approach for the assessment of change and detection of response shifts, we would like to make the following recommendations: (1) take into account substantive considerations when making decisions on model fit evaluation (e.g., using theory to establish an appropriate measurement model, or substantive knowledge and practical interpretation of response shift effects, in addition to relying on model fit tests or indices to guide the (re)specification of the model). (2) Use several tests and indices of (differences in) model fit to prevent overly simplified decision rules based on single criteria. Instead, find support for the robustness of results warranted by multiple criteria. (3) Keep in mind that some fit indices are more appropriate in certain circumstances than others (e.g., specifically developed to take into account model parsimony). (4) Try to interpret and possibly explain detected response shift effects to substantiate the linkage between detected response shift and experience of response shift by patients (e.g., using substantive knowledge and practical interpretation, or direct measures of possible explanatory variables). It is our aim that these recommendations will help to stimulate the appropriate application and interpretation of SEM methodology for the investigation of response shift and assessment of change.

### Directions for Future Research

The assessment of patient-reported outcomes, including HRQL, are becoming standard part of clinical practice (Osoba, 2011). In order to understand patterns of change in patients' HRQL, the investigation of response shift will thus become more important. It is anticipated that researchers will be increasingly involved in the analyses of complex data structures, such as data from extensive longitudinal assessments (e.g., long-term follow-up assessments or momentary assessments using mobile phone devices), from planned missing data designs (e.g., measures in groups of patients with different time-lags between measurements), or from combinations of separate clinical trials (e.g., including different groups of patients from different treatment centers or with different treatment regimens). These developments can benefit from advanced SEM methodology (e.g., exploratory models, multi-level models). At the same time, accessibility of (advanced) SEM methodology is facilitated thanks to the availability of (free) software programs (e.g., Lavaan; Rosseel, 2012), textbooks and tutorials that are specifically aimed at practical applications (e.g., Kline, 2015; Schumacker & Lomax, 2016; Rosseel, 2016), including those in the field of HRQL (e.g., Fayers & Machin, 2016).

Therefore, there is great potential for future studies to focus on the development and explanation of response shift detection methods; especially for increasingly complex data structures. Appropriate applications of response shift detection methods could be further enhanced by focusing on the comparison of different methods for response shift detection (e.g., providing guidelines on which method to use under what circumstances), the cross-validation of results (e.g., increasing generalizability of findings), and the facilitation of sharing data and syntaxes (e.g., improving transparency and the possibility of replication of results by other researchers). Such research endeavors may enhance the scientific stringency of applications and interpretation of SEM methodology for the detection of response shift effects, and help to advance research into HRQL.

To improve our understanding of change in HRQL outcomes, future research could focus on the interpretation and explanation of response shift effects. The combination of qualitative and quantitative research may be used to provide the necessary substantive linkages between detected response shift effects and response shift as experienced by patients. In addition, direct measurements of catalysts, antecedents, or mechanisms could be used to investigate possible predictors of different types of change. Moreover, the interpretation of the clinical significance of change, including response shift effects, may be an important area for future research; for example by improving the standard reporting of effect-size indices of change. The interpretation and explanation of response shift will help to translate the findings of response shift research to clinicians and patients, and will ultimately help to adopt more effective treatments and thus enhance patients' HRQL.