Unbiased measurement of health-related quality-of-life
King, B.L.

Citation for published version (APA):
King, B. L. (2011). Unbiased measurement of health-related quality-of-life
Introduction: Unbiased Measurement of Health-Related Quality-of-Life
By measurement to knowledge [door meten tot weten] I should like to write as a motto above the entrance to every physics laboratory.

— Heike Kamerlingh Onnes

The above quote captures a common ideal not only in physics, but more generally in science regarding measurement. Knowledge, no matter the discipline, can only be obtained when we have a good measure. Measuring the distance from the edge of my desk to the door will, as long as I do not move my desk and ceteris paribus (all things remaining equal), always be 3.55 meters, no matter on which day I measure the distance. If a person’s blood pressure is measured on two different days, we do not necessarily expect the exact same value; however, we believe that we are measuring the pressure of the circulating blood against the wall of the blood vessels, which is produced primarily by the contractions of the heart [1]. If we measure someone’s self-assessed health-related quality-of-life (HRQoL) it becomes less clear what we are measuring, as the meaning of the construct of HRQoL may be subject to change, even within the same person. For example, what HRQoL means to a patient after a diagnosis of cancer may be different from what HRQoL means to the same patient after receiving chemotherapy.

**Measurement invariance: What is it?**

More than half a century ago, Campbell [2] discussed how extraneous factors in the analysis of change could confound the assessment of change in the outcome of interest. Among the possibilities discussed was the idea that respondents could change their frame of reference leading to instrument decay. This was put more formally by Cronbach and Furby [3], who noted that all analysis of change hinges on the assumption that there is a common metric of the latent variable across measurement occasions. Since this time different fields have developed theories that encompass these ideas and attempt to explain, test and control for confounding due to different metrics of the latent variable in the
analysis of change. The theories can be divided into two groups, psychometric and substantive.

*Psychometric Theory*

In 1989 Mellenbergh [4] formally presented a general formula that can be applied to the situation Campbell [2] and Cronbach and Furby [3] described and that represents the field of measurement invariance. Measurement invariance is defined as:

\[ f_1(X|A = a, V = v) = f_2(X|A = a) \]

where \( X \) is an observed item, \( A \) the attribute we would like to measure and \( V \) any variable that has the potential to violate the relationship between \( X \) and \( A \) and thus lead to measurement bias. Functions \( f_1 \) and \( f_2 \) are both conditional distributions, \( f_1 \) being the conditional distribution of \( X \) given \( a \) and \( v \), and \( f_2 \) the conditional distribution of \( X \) given only \( a \). If the conditional independence does not hold, \( (f_1 \neq f_2) \), then the measurement of \( A \) by \( X \) is said to be biased by \( V \) and the assumption of measurement invariance is violated. What \( V \) represents is related to the research question. In this thesis \( V \) primarily represents time, but is also used to represent group membership.

Mellenbergh’s [4] work has provided a statistical framework for testing invariance, which has been further developed by a number of researchers, notably by Meredith [5], who applied the above formula to a multi-group confirmatory factor analysis (CFA). This application has been extensively utilized in diverse fields of research, where group membership has been defined by gender, race, age, disease, employment status just to name a few. In an extensive review of this literature, Vandenberg and Lance [6] reported that there are three tests of measurement invariance and thus the detection of measurement bias that are of primary importance. While the terminologies that are used differ, the three tests are configural invariance (same pattern of free and fixed elements in the factor loading matrix), metric invariance (equal factor loadings) and scalar invariance (equal factor loadings and equal intercepts) [6]. Until recently, most work was focused on multiple-group invariance testing, however these tests can be easily applied to longitudinal data [7-9]; and can be considered crucial for the
comparison of common factor means, either by group membership or across time.

Substantive Theories

In parallel to the theory of measurement invariance, substantive researchers developed their own taxonomies for describing a shifting metric for common attributes specific to longitudinal research. Two prominent taxonomies, developed at the same time, are from the fields of organization/management research [10] and educational research [11]. The first theory is known as alpha, beta, gamma change, where alpha change is defined as true change in the attribute of interest; beta change as change due to respondents’ recalibration of the measurement scale; and gamma change as the change in the respondents’ understanding of the attribute of interest [10]. From the field of educational research, Howard et al. [11] coined the term response shift, which refers to change experienced by respondents in their internal standard of measurement from one measurement occasion to the next, which is similar to beta change. This change in measurement is the result of some (educational) intervention that alters the way the participant regards the construct of interest.

The taxonomy adopted in the current thesis was developed by Sprangers and Schwartz [12], which combines elements of both Golembieski et al. [10] and Howard et al. [11] and is applied to the field of HRQoL research. Sprangers and Schwartz’s work was in response to paradoxical or ambiguous findings in HRQoL research, where patients facing a life threatening illness, report similar or better HRQoL in comparison to healthy individuals. Such findings might be the result of recalibration, reprioritization, or reconceptualization; which occur as a result of health state changes [12]. Recalibration is similar to beta change and reconceptualization to gamma change. Reprioritization is defined as a change in respondents’ values in that different domains may become more or less important over time.

To marry the terminologies used in psychometerics and HRQoL research, it can generally be said that when measurement invariance is violated with respect to time, this can be considered as a special case of measurement bias, that is often referred to as ‘response shift’. More specifically, a violation of
configural factor invariance corresponds to reconceptualization, a violation of metric measurement invariance to re priorititation and a violation of scalar measurement invariance to recalibration. In this introduction we will use the terminology associated with response shift. However, throughout the thesis, terminologies are mixed because some chapters are intended for HRQoL researchers and others for psychometricians.

**Measurement invariance: Why investigate it.**

To measure patient reported outcomes, such as HRQoL, we rely on the use of self-report questionnaires. Observed scores on self-report questionnaires are particularly susceptible to bias or response shift. In HRQoL research this susceptibility to bias or response shift is hypothesized to be related to a catalyst [12]. Examples of such catalysts are a potentially life threatening disease or invasive treatment. Response shift can also be related to an intervention to change patient behavior or the mere passage of time. Regardless, if patients change their frame of reference between consecutive measurement points our conclusions regarding change in HRQoL may be wrong. This is because we may not be entirely capturing true change in HRQoL but rather true change and systematic bias as a result of response shift [9].

To provide an example of how the three conceptualizations of response shift can occur, imagine a lung cancer patient, Lisa, who has just been diagnosed and has agreed to participate in a study investigating HRQoL. When filling in the questionnaire, Lisa comes upon the item “How much of the time during the past 4 weeks did you feel tired?” Lisa has a full time job and two young children and feels tired a lot, so she responds with “Most of the time = 2”. Lisa then undergoes chemotherapy and after her last treatment cycle she feels exhausted, even after sleeping for eight hours. At that time she completes the same questionnaire and again responds with “Most of the time, = 2” to the fatigue item. However, the value of “2” does not represent the same level of fatigue she experienced prior to chemotherapy, therefore Lisa has recalibrated the measurement scale. It is also possible that Lisa has reconceptualized what being tired means. For example, before chemotherapy Lisa would say she was very tired after a full day of work.
and taking care of one of her sick children. It is also possible that while Lisa feels tired most of the time, this is now less important to her than before as she enjoys the emotional support provided by her family and being able to spend time with them. Therefore she has reprioritized fatigue so that it is now less important to her than social functioning, whereas prior to treatment these two domains may have been equally important. If we only look at the response Lisa provided to the item, we would conclude that Lisa’s assessment of fatigue had not changed.

To prevent such erroneous conclusions, it is important to test measurement invariance over measurement occasions to ensure that a response shift has not occurred, and to test that no measurement bias is present in respect to additional patient and disease characteristics. In testing these assumptions and accounting for any response shift or measurement bias identified the conclusions regarding change in the construct of interest are valid. This is true for any self-report measure, and not only important for HRQoL measures.

**Measurement Invariance: How to Investigate it.**

Various methods have been proposed for investigating bias and response shift. Golembieski, et al. [10] proposed using factor analysis, a statistical approach; whereas Howard et al. [11] proposed using the then-test, a design approach. In general, design approaches require additional measures to assess whether the metric of the construct of interest has changed and whether reprioritization and reconceptualization has occurred. Statistical approaches on the other hand, do not necessarily have this requirement as the choice of the method can be made post hoc based on the available data [13]. The work of Jöreskog and Sörbom [14;15] inspired both psychometricians [5] and substantive researchers [16] to develop statistical procedures for investigating bias and response shift. Oort [9;17] using structural equation modeling (SEM) brought both fields together using the framework of Sprangers and Schwartz [12].

A number of other advanced statistical methods are used in bias and response shift evaluation. Recently a head-to-head comparison of some of these methods was conducted using the same dataset to investigate response shift in two HRQoL measures. Methods that were compared included SEM [18], latent
trajectory analysis [19], and classification and regression tree modeling [20]. Regardless of the method, very few instances of response shift were identified. The response shifts that were identified did differ between methods; however these results may be more reflective of how the different methods operationalized the observed variables. These papers were presented at the 17th Annual Conference of the International Society for Quality of Life Research [21] and will be published in an upcoming issue of Quality of Life Research.

In this thesis, we focus on the approach suggested by Oort [9;17]. In using his procedure, measurement bias and response shift can be detected, accounted for and true change in HRQoL can be assessed. This requires a questionnaire with multiple items, aimed at measuring the attribute of interest (e.g., HRQoL), which is administered on at least two measurement occasions. To identify response shift a hierarchical four step procedure is proposed by Oort. The four steps are: 1) Establishing a measurement model; 2) Overall testing of response shift; 3) Response shift detection; and 4) True change assessment. Once a valid measurement model has been identified for the questionnaire items, equality constraints are placed on the parameters of interest, and the tenability of these constraints are investigated. When the constraints are not tenable this is strong indication of bias [9;17]. In this thesis we focus primarily on factor loadings and intercepts. Non-tenable equality constraints associated with the factor loadings represent either reconceptualization (zero to non-zero estimates) or reprioritization (strength of estimate changing over measurement occasions). Non-tenable equality constraints associated with the intercepts represent recalibration.

**Measurement Invariance: What to do if Bias is Identified**

When measurement invariance has been violated and thus bias or response shift are present, true change in HRQoL can still be investigated. However, before doing so the bias must be accounted for. If a biased parameter is identified in Step 3 of Oort’s [9;17] procedure, then the equality constraints placed on the parameter(s) are not tenable and therefore removed. When equality constraints are removed, we have a scale with partial measurement
invariance. While it is ideal when comparing common factor means to have full measurement invariance and therefore no bias or response shift, this requirement may be too restrictive. Therefore, Byrnes, et al. [22] were among the first to argue that full measurement invariance is not a necessary condition for common factor mean comparisons, but rather a good starting point.

After establishing the invariance or partial invariance of factor loadings and intercepts then Step 4 of Oort’s procedure can be carried out where the effect sizes of the response shift, observed change and true change can be calculated using the partitioning formula [9]. These effect sizes can be interpreted in terms of Cohen’s effect size $d$ [23]. In addition to investigating the size of response shift associated with biased parameters, the common factor mean can now be investigated to determine if HRQoL has changed between measurement occasions. Once bias and response shift have been accounted for, any change in HRQoL is valid and can be interpreted with confidence.

**Measurement Invariance: The Impact on Substantive Conclusions**

If we return to the example of Lisa and her evaluation of fatigue, one can see that her internal standards, prioritization and conceptualization when assessing fatigue have changed. What is important from a measurement perspective is that the assumption of measurement invariance has been violated and the assessment of change has been compromised. In this thesis, this is our primary focus, for example we aim to account for response shift and measurement bias so that we can measure true change in for example fatigue associated with HRQoL.

While bias and response shift detection has become more popular, more work is needed to better understand the phenomenon in diverse patient populations who experience different catalysts that could result in a health state change. Second to this, the methods and procedures used to detect bias and response shift are rarely consistently applied and have thus far led to mixed results. Regardless, when bias or response shift is present in the data, it has the potential to substantively affect the conclusions drawn [9] and therefore requires consideration.
Outline of Thesis

When measurement invariance is violated and response shift or measurement bias is present, any observed change may reflect a true change in HRQoL or a change that is associated with the bias. While there is a general consensus that this is problematic, we currently do not know the best procedure for testing measurement invariance, which factors will affect measurement bias and response shift and under what circumstances these factors play a role. Therefore, the general objective of this thesis is to investigate measurement invariance in existing sets of HRQoL data in diverse patient populations, to account for measurement biases and response shifts, and assess true effects on unbiased HRQoL. This will be achieved by using SEM. In applying the procedure outlined by Oort [9] to the data used in this thesis we anticipate that methodological problems will arise. As a result, our secondary aim is to address some of these problems in two methodological papers where we will use empirical examples to illustrate the problem. As a result this thesis is divided into two sections; applied papers and methodological papers.
### Table 1. Objectives of thesis

<table>
<thead>
<tr>
<th>Objectives</th>
<th>Patient Population</th>
<th>Instrument</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Chapter 2</strong></td>
<td>Breast, lung, pancreatic and esophageal cancer patients (n = 202)</td>
<td>SF-36 [24]</td>
</tr>
<tr>
<td>To illustrate how two perspectives of response shift are related; the measurement perspective and conceptual perspective</td>
<td></td>
<td></td>
</tr>
<tr>
<td>To investigate bias and response shift associated with different patient characteristics: sex, age, cancer site, health status, optimism, and upward comparison</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Chapter 3</strong></td>
<td>Multiple Sclerosis patients (Study 1, n = 1,552; Study 2, n = 1,767)</td>
<td>SF-12 [25] and the Disability Scales [26]</td>
</tr>
</tbody>
</table>
### Methods Chapters

<table>
<thead>
<tr>
<th>Chapter 4</th>
<th>To investigate measurement bias in a cross-sectional sample using restricted factor analysis</th>
<th>Heterogeneous cancer patients (n = 155)</th>
<th>EORTC QLQ-C30 [27]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>To investigate bias associated with different patient characteristics: sex, age, previous treatment for cancer, and information regarding treatment preferences</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chapter 5</td>
<td>To investigate bias and response shift in a longitudinal (3 measurement occasions), multi-group (two treatment groups) analysis</td>
<td>Patients who screened positive for depression, anxiety, and/or at-risk drinking (n = 1,198)</td>
<td>SAMHSA Mental Health and Alcohol Abuse Stigma Assessment</td>
</tr>
<tr>
<td></td>
<td>To investigate whether the wording of items has a direct effect on observed item scores when they should not</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>To investigate bias associated with different patient characteristics: sex, age, education, race, Physical and Mental HRQoL, mental health symptoms, and at-risk drinking problem</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chapter 6</td>
<td>To illustrate a refined procedure for detecting bias and response shift</td>
<td>HIV/AIDS patients (n = 403)</td>
<td>MOS-HIV [28]</td>
</tr>
<tr>
<td></td>
<td>To present this procedure in a series of hierarchical steps that are easy to follow</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
To address the issue of chance findings
To address the problem of the constraint interaction and how to avoid it

Chapter 7
• To illustrate two different strategies for detecting bias and response shift
• To highlight advantages and disadvantages of using different strategies in detecting bias
• To address the issue of chance findings in relation to single and multiple parameter tests

Lung cancer patients (n = 216) EORTC QLQ-C30 [27] and EORTC-LC13 [29]
References


