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
General Discussion
The largely objective basis of statistical algorithms does not remove the need for human judgment in their implementation

— Ken A. Bollen, 2000

The general objectives of this thesis were to investigate measurement invariance in existing sets of HRQoL data in diverse patient populations, account for measurement biases and response shifts, and assess true effects on unbiased health-related quality-of-life (HRQoL). In applying the procedure to detect response shift outlined by Oort [1;2] to the data used in this thesis we anticipated that methodological problems would arise. As a result, our secondary aim was to address some of these problems in two methodological papers that used empirical examples to illustrate the problem.

In applying structural equation modeling (SEM) to investigate measurement invariance, we were able to meet our first objective as the samples were diverse, bias and response shift were detected and accounted for, and we were able to assess true change in the measurement of HRQoL. Table 1 provides a summary of these results. In regards to the second objective, we proposed modifications to Oort’s procedure [1;3] for testing measurement invariance and illustrated this new procedure with data from HIV/AIDS patients. This procedure was further extended (by including inspection of observed parameter changes) and compared to another procedure (relying on modification indices and expected parameter changes) for investigating measurement invariance. In proposing this refined procedure, we were able to better guard against chance findings.

Diverse patient populations

Until now, few studies have applied the same procedure to investigate measurement bias and response shift across a diverse group of patient populations. The diversity of patient populations used to investigate measurement invariance is both a strength and weakness of this thesis. It is a
strength because this thesis provides a broad overview of biases and response shifts that may be identified in different patient populations, with both acute (e.g., cancer) and chronic (e.g., multiple sclerosis) conditions. It is a weakness because we are not able to easily compare the results across data sets.

**Bias and response shift**

In all four of the applied papers (Chapters 2 - 5) and both of the methods papers (Chapters 6 and 7), either bias or response shift was identified. Most frequently, bias and response shifts were associated with observed variables that utilized items with more general wording (e.g., SF-36 [4], item 11b – I am as healthy as anybody I know). These observed variables are particularly susceptible to bias because they are unspecific questions which leave room for respondents to attach their own meaning, allowing for patient characteristics to affect the responses provided.

In Chapters 3 and 6 the biases and response shifts identified were small in magnitude and sometimes difficult to interpret. In both of these samples, no clear catalyst (for example, starting treatment) occurred for all respondents. Such a catalyst may be required so that a significant change in health state occurs, which in turn has the potential to lead to a meaningful and interpretable response shift [5].

The additional exogenous variables associated with measurement bias and response shifts varied greatly across studies. However, in three out of five papers (Chapters 2, 3 and 4) where age was included as an exogenous variable, bias with respect to age was detected. Improving our understanding of the relationship between age, HRQoL and measurement bias is an area in need of further research. It is already well documented [6;7] that perceptions of aging influence health outcomes. Interesting questions that require further research include: are the effects truly age effects or are they cohort effects. This question requires studying different age groups over a significant period of time to assess whether the effects are consistent for age groups or cohort groups. Another important research question is whether the inclusion of age specific HRQoL questionnaires (e.g., EORTC QLQ-ELD15 [8]) would be beneficial in avoiding bias.
with respect to age. If so, it is also important to understand how these could be incorporated into studies where a broad age range is studied. Alternatively one might want to include, for example, the Expectations Regarding Aging Survey [9] and test for bias with respect to scores on this scale.

**True effects**

After accounting for bias and response shift, change in the common factor means was investigated. Significant change in the common factors was identified in most of the studies. In Chapter 7 where two procedures for detecting bias and response shift were compared, the conclusions from the two procedures differed regarding change in the common factor means after accounting for bias. We concluded that there was no significant change in Physical HRQoL over time when using the procedure relying on modification indices and expected parameter changes. Whereas, we concluded that there was a small significant decrease in Physical HRQoL when using our Three Step procedure (relying on global tests and observed parameter changes). This result can be explained by the different response shifts that were accounted for in each procedure, which was primarily as a result of testing single parameters or testing multiple parameters simultaneously.

The impact of response shift and true change on observed change varied among chapters. Using Oort’s [1] partitioning formula we were able to determine the impact of response shift and true change in terms of Cohen’s $d$ effect sizes [10]. All effect sizes were small in magnitude. However, by distinguishing response shift from true change, we found that the conclusions we would have arrived at changed. For example, in Chapter 3, we would have concluded that the item measuring sensory disability had changed for stable multiple sclerosis patients (Study 1). However, most of this change was attributable to recalibration response shift. Therefore after response shift was accounted for, there was only a very small true change in sensory disability for stable patients.
Methodological considerations: procedures and chance findings

As we applied the procedure originally outlined by Oort [1;3] it was discovered that chance findings may arise and that by refining the procedure, we could reduce those chance findings. In Chapter 6 we refined the procedure to guard against chance findings in three ways: a) by directly testing specific hypotheses, b) by using global tests to simultaneously consider the invariance of multiple parameters (both factor loadings and intercepts), and c) by conducting all tests at Bonferroni adjusted levels of significance [11].

Overall this new procedure was found to reduce chance findings and also circumvent the constraint interaction. However, when the procedure was applied to large samples it became clear that the global tests were significant when the observed difference in parameter estimates was small, thus chance findings were not substantively reduced. To overcome this, we proposed investigating the observed parameter change (OPC). We believe that when the global tests and OPCs are used in conjunction, chance findings will be further reduced, and a more accurate assessment of change obtained.

Advantages and disadvantages of the Three Step procedure

The procedure presented in this thesis consists of three broad steps: 1) establishing a measurement model, 2) testing measurement invariance across measurement occasions, and 3) testing measurement invariance with respect to exogenous variables. The advantages of the procedure using SEM are that: a) it is sufficiently general to be applied in almost any setting and under different circumstances; b) it retains construct validity; and c) it is rigorous, and guards against chance findings. With this rigor comes the disadvantage that this procedure is time intensive to carry out. This is because we test all alternative models, whereas other procedures only test models for which violations of invariance are plausible, as for example, indicated by modification indices or standardized residuals. To illustrate this difference imagine there was one common factor of HRQoL, five observed variables and one instance of response shift. Using our procedure nine models would be fit; relying on modification indices, for example, only two models would be fit. Even in this simple example,
the difference is quite large. This disadvantage is problematic, as testing the assumption of measurement invariance will not always be the primary outcome of interest in a study. Applied researchers who want to evaluate this assumption may not carry out the procedure suggested in this thesis for this reason. Despite this disadvantage, by considering each alternative model, the researcher will gain insight into the substantive effects of freeing the equality constraints, and therefore gain increased understanding of his or her model and reduce the effects of relying on a data driven procedure. Finally, it is possible that the cut-off values we chose for the OPCs, which are based on Cohen’s $d$ small effect sizes [10], may not always be ideal under different data circumstances. However, the proposed cutoff values can be changed and a more appropriate cutoff value can be chosen.

**Improving the Three Step procedure**

To increase and further our understanding of the three step procedure, simulation research is required. This would allow us to test the performance of our procedure under commonly arising circumstances in HRQoL research (e.g., skewed distributions, small sample size). We can then evaluate when the procedure is and is not able to reduce chance findings. If the simulation work would lead to suggesting different guidelines for different sample sizes, for example, then the next step would be to test our procedure against the new recommendations. To ensure comparability of results, the same empirical data would be used to investigate whether the same or different results are identified. Such research is warranted as it would increase our confidence in our conclusions regarding change in HRQoL data, or any construct that has been measured with self-report questionnaires.

**Final Remarks**

In general, the findings presented in this thesis highlight the importance of testing the assumption of measurement invariance. As all studies exhibited some level of measurement bias or response shift, these phenomena should not be overlooked. We relied solely on SEM to detect measurement bias and
response shift as we believe it to be an effective method to investigate measurement invariance post study design. This is due to the flexibility it offers researchers as the analyses can be tailored to most study designs. In addition to this, it may be the most accessible technique for social scientists, as it offers a framework that is similar to factor analysis, which most social scientists have encountered at some point during their careers. In this thesis we demonstrated some of this flexibility and hope that the proposed procedure offers a relatively user friendly approach to testing measurement invariance in self-reported measures.
Table 1. Summary of results

<table>
<thead>
<tr>
<th>Chapter Population</th>
<th>Instrument</th>
<th>Response Shift</th>
<th>Measurement Bias</th>
<th>True Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapter 2: Breast, lung,</td>
<td>SF-36 [4]</td>
<td>• Bodily pain (reconceptualization) – response affected by both Physical &amp; Mental</td>
<td>• Older age → worse Physical Functioning (T1 &amp; T2)</td>
<td>• Significant decrease in Physical HRQoL</td>
</tr>
<tr>
<td>pancreatic and esophageal</td>
<td></td>
<td>Functioning at T2</td>
<td>• Female sex → better General Health (T1)</td>
<td>• Significant increase in Mental HRQoL</td>
</tr>
<tr>
<td>cancer (n = 202)</td>
<td></td>
<td>• General health (reconceptualization) – response affected by both Physical &amp;</td>
<td>• More optimism → better General Health (T2)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mental HRQoL at T2</td>
<td>• More upward comparison → worse General Health (T2)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• More optimism → better Mental HRQoL (T1 &amp; T2 – explanation bias)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• More upward comparison → worse Mental HRQoL (T1 &amp; T2 – explanation bias)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Significant decrease in Physical HRQoL</td>
<td></td>
</tr>
<tr>
<td>Chapter 3: Multiple</td>
<td>The Disability Scales</td>
<td>• Study 1: Sensory disability (recalibration) – less inclined to report high</td>
<td>• Study 2: Older age → better Mental Health (T1, T2, T3)</td>
<td>• Study 1: Overall deterioration in Visible Disability</td>
</tr>
<tr>
<td>Sclerosis (Study 1,</td>
<td>(Study 1) [12]; SF-12</td>
<td>sensory disability at T2 &amp; T3</td>
<td>• Study 2: Older age → better Vitality (T1, T2, T3)</td>
<td>• Study 1: No change in Internal Disability</td>
</tr>
<tr>
<td>n = 1,552; Study 2, n =</td>
<td>(Study 2)</td>
<td>• Study 2: Role Emotion (recalibration) – less inclined</td>
<td>• Study 2: No change in Internal Disability</td>
<td>• Study 2: No change in</td>
</tr>
<tr>
<td>1,767)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chapter 4</td>
<td>Heterogeneous cancer (n = 155)</td>
<td>EORTC QLQ-C30 [14]</td>
<td>Cross sectional data – not applicable</td>
<td>Either Physical or Mental HRQoL</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Chapter 5</td>
<td>Screened positive for depression, anxiety, and/or at-risk drinking (n = 1,198)</td>
<td>SAMHSA Mental Health and Alcohol Abuse Stigma Assessment</td>
<td>Item 4: Would it be difficult for you to start mental health or alcohol abuse treatment if other people knew that you were going to be in treatment? (recalibration) – more inclined to report high on item 4 at T2 &amp; T3</td>
<td>Older age → worse Physical Functioning</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Female sex → more stigma on Item 3 (incremental change T1, T2, T3)</td>
<td>Previous Treatment → worse Emotional Functioning</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Better mental HRQoL → less stigma on Item 3 (incremental change T1, T2, T3)</td>
<td>Better mental HRQoL → more comfort on Item 6: How comfortable would you be talking about your mental health or alcohol abuse problems with a counselor or mental health professional? (T1, T2, T3)</td>
</tr>
</tbody>
</table>
### Chapter 6: HIV/AIDS (n = 403)

- **MOS-HIV** [15]
  - Health Distress (recalibration) - more inclined to report lower health distress at T1
  - Emotional Functioning (recalibration) - more inclined to report lower emotional functioning at T2, T3 & T4
  - Lower CD4-cell count → better Emotional Functioning (T1, T2, T3, T4)
  - Lower CD4-cell count → worse Role Functioning (T1) better Role Functioning (T2, T3 & T4)
- Significant increase in Physical and Mental HRQoL

### Chapter 7: Lung cancer (n = 216)

- **EORTC QLQ-C30** [14] and **EORTC-LC13** [16]
  - Procedure 1: Emotional Functioning (recalibration) - more inclined to report lower emotional functioning at T1
  - Procedure 2: Dyspnea (recalibration) - more inclined to report worse dyspnea at T1
  - Procedure 2: Nausea (recalibration) - more inclined to report worse nausea at T2
- No additional exogenous variables included in model – not applicable
- Procedure 1: No significant change in Physical or Mental HRQoL
- Procedure 2: Significant change in Physical and not in Mental HRQoL

---

**Note:** T refers to the measurement occasion
References


