AStA Adv Stat Anal (2010) 94: 139–156
DOI 10.1007/s10182-010-0129-y

ORIGINAL PAPER

Using structural equation modelling to detect measurement bias and response shift in longitudinal data

B.L. King-Kallimanis · F.J. Oort · G.J.A. Garst

Received: 24 September 2009 / Accepted: 30 January 2010 / Published online: 26 May 2010

Abstract We propose a three step procedure to investigate measurement bias and response shift, a special case of measurement bias in longitudinal data. Structural equation modelling is used in each of the three steps, which can be described as (1) establishing a measurement model using confirmatory factor analysis, (2) detecting measurement bias by testing the equivalence of model parameters across measurement occasions, (3) detecting measurement bias with respect to additional exogenous variables by testing their direct effects on the indicator variables. The resulting model can be used to investigate true change in the attributes of interest, by testing changes in common factor means. Solutions for the issue of constraint interaction and for chance capitalisation in model specification searches are discussed as part of the procedure. The procedure is illustrated by applying it to longitudinal health-related quality-of-life data of HIV/AIDS patients, collected at four semi-annual measurement occasions.

Keywords Measurement bias · Response shift · Longitudinal data · Structural equation modelling · Factor analysis

B.L. King-Kallimanis (✉)
Department of Medical Psychology, Academic Medical Centre, University of Amsterdam, Amsterdam, The Netherlands
e-mail: B.L.Kallimanis-King@amc.uva.nl

F.J. Oort · G.J.A. Garst
Department of Education, University of Amsterdam, Amsterdam, The Netherlands

F.J. Oort
e-mail: F.J.Oort@uva.nl

G.J.A. Garst
e-mail: G.J.A.Garst@uva.nl

1 Introduction

In longitudinal research, the aim generally is to assess and explain change in the research subjects' attributes. In behavioural research, the attributes of interest are often measured subjectively, through self-report questionnaires. When analysing change as the result of some event, for example an intervention, measurement invariance is assumed. However, if the assumption of measurement invariance is not tested, we cannot be sure whether the change in observed test scores fully represents true change in the attribute of interest, or also change in the response behaviour of the respondent. If this assumption is violated and measurement bias is present, the assessment of change is compromised. If researchers overlook this assumption then they risk the validity of their substantive interpretations.

Measurement invariance is defined as

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

where X is a set of observed variables (e.g. items or scales of a questionnaire), T represents the attributes of interest (the theoretical constructs the questionnaire is designed to measure) that are measured by X, and V represents variables that could potentially violate measurement invariance (e.g. any other attribute than those represented by T, or experimental condition, or time, etc.). Function f_1 is the conditional distribution function of X given values t and v, and f_2 is the conditional distribution function of X given t. In other words, measurement invariance implies that respondents with equal standings on the attributes of interest T have equal expected values of the response variables X and no other variables V systematically affect response variables X. If the conditional independence does not hold, that is, if $f_1 \neq f_2$, then the measurement of T by X is said to be biased by V and the assumption of measurement invariance is violated. The definition, as introduced by Mellenbergh (1989) is very general as it defines measurement invariance as statistical independence (not just linear independence), and variables X, T, and V can be of any measurement level, continuous or discrete, observed or latent.

Meredith (1993) used Mellenbergh's definition to define weak (linear) measurement invariance and applied it to multigroup confirmatory factor analysis (CFA). In this application, X generally are observed continuous indicators, T are latent continuous variables (common factors), and V are observed discrete variables, defining some group membership. This application of measurement invariance analysis in multigroup CFA is well known and has been reviewed by Vandenberg and Lance (2000) and Schmitt and Kuljanin (2008). Terminologies vary, but we may distinguish configural factorial invariance (same patterns of fixed and free elements in factor loading matrices), weak factorial invariance (same factor loadings), strong factorial invariance (same factor loadings and intercepts), and strict factorial invariance (same factor loadings, intercepts, and residual variances). Similar tests for configural, weak, strong, and strict factorial invariance can and have been conducted with longitudinal data (Sayer and Cumsille 2001; Meredith and Horn 2001; Oort 2005a). The measurement invariance definition then still applies, with V representing an index for the time of the measurement occasion (Oort 1991, 2005b). When

this assumption of measurement invariance is violated with respect to time, we can consider this a special case of measurement bias, often referred to as "response shift".

The term response shift was coined by Howard and colleagues who conducted research on educational training interventions (Howard et al. 1979). They defined response shift in terms of changes in internal standards of measurement, obfuscating true change in the attribute of interest. Golembiewski et al. (1976), researching organizational change, described three types of change, which they labelled alpha, beta, and gamma change. Alpha change refers to objective change (or true change), beta change is a change in the meaning that respondents attach to the labels of response scale points, and gamma change refers to a change in the respondents' understanding of item content, that is, the meaning of the wording of the items in the questionnaire. Response shift seems to encompass both beta and gamma change. In the field of medical psychology, Sprangers and Schwartz (1999) distinguish three types of response shift. Recalibration response shift is a change in the respondent's internal standards of measurement, reprioritization response shift is a change in the respondent's values, and reconceptualization response shift is a redefinition of the target construct. According to Oort (2005a), recalibration response shift (or beta change) violates intercept invariance, reprioritization response shift (or gamma change) violates factor loading invariance, and reconceptualization response shift (or gamma change as well) violates the invariance of factor loading patterns.

The purpose of the present paper is to show how structural equation modelling (SEM) can be used to investigate measurement bias and response shift in longitudinal data, by applying SEM to health-related quality-of-life data collected from HIV/AIDS patients over four semi-annual measurement occasions. Measurement bias detection in longitudinal data from four time points (instead of two) highlights the problem of chance findings. Here we will propose global tests at Bonferroni adjusted levels of significance to reduce the number of chance findings. The global tests also solve the problem with so-called interaction constraints (i.e. when arbitrary scaling choices affect test results). In addition to the consideration of these two issues, the example also serves to illustrate how response shift is related to measurement bias in two ways: response shift as measurement bias with respect to time, and response shift as measurement bias with respect to exogenous variables.

2 Methods

The goal of the procedure presented below is to detect and account for measurement bias and response shift in longitudinal data, in order to validly assess "true" change in the attributes of interest.

The procedure has three steps. The goal of the first step is to find a confirmatory factor model that constitutes an appropriate measurement model with good fit and clear interpretation. In the second step, the longitudinal factor model is used to detect measurement bias by testing the invariance of factor loadings and intercepts across measurement occasions. In the third step, we add exogenous variables to the measurement model, and detect measurement bias by testing direct effects of these variables on the observed indicator variables. After detecting and accounting for apparent bias,

change in the attributes of interest can be evaluated by assessing the differences in the common factor means.

2.1 Measurement bias and response shift

In the second step, the measurement bias that is investigated is covered by the measurement invariance definition if we substitute an index of time or measurement occasion for V. We therefore consider any bias detected in this step to be response shift (Oort 1991, 2005b). In the third step, measurement bias is investigated with respect to exogenous variables V that are possibly related to the attributes of interest and that are suspect to induce bias in the observed indicator variables. If the size of such measurement bias varies across measurement occasions, then such measurement bias may be considered as response shift as well (Oort 2005b; Oort et al. 2009).

2.2 Structural equation modelling

The three-step procedure relies on fitting series of structural equation models and comparing their fit. The maximum likelihood estimation method yields a chi-square measure of overall goodness-of-fit that constitutes a test of the equivalence of the model implied means, variances, and covariances and the observed means, variances, and covariances. As this test of exact fit is very sensitive to small deviations, we additionally consider the root mean square error of approximation (RMSEA) as an index of approximate fit. According to a generally accepted rule of thumb, RMSEA values smaller than 0.05 suggest close fit and values smaller than 0.08 suggest satisfactory fit (Browne and Cudeck 1992).

Chi-square difference tests will be used to assess the appropriateness and significance of changes made to the model as a result of testing measurement invariance. It is important to note that these difference tests can also be used to compare the fit of nested models when neither model shows good fit, if only we can assume equivalence of the non-centrality parameters of the associated non-central chi-square distributions (Steiger et al. 1985). The chi-square difference tests can be complemented by calculating the difference in the expected cross-validation index (ECVI) for nested models. The ECVI is an information criterion that is linearly related to the Akaike's criterion (Browne and Cudeck 1992). If a 90% confidence interval for an ECVI difference between nested models includes zero then the fit of the two models is considered to be essentially equivalent (ibidem).

2.3 Circumventing constraint interaction

The results of the chi-square difference tests should not depend on the arbitrary choice of scale and origin of the common factors. In our procedure, we will choose to impose scale and origin by fixing the variance and mean of the common factors of the first measurement occasion. The common factors of the other measurement occasions are then given scales and origins through invariance constraints on factor loadings and intercepts. However, the detection of measurement bias involves testing and possible

removal of these constraints. If we would choose, as customary, to investigate the invariance of factor loadings first and intercepts second, then the removal of one or more factor loading constraints would interact with the chi-square difference tests of subsequent intercept constraints. Such constraint interaction is caused by the fact that the model implied means (μ) of the observed variables are a function of both intercepts (τ) and the product of factor loadings (Λ) and common factor means (κ), $\mu = \tau + \Lambda\kappa$. As a result, an across occasion difference in Λ affects the across occasion difference in μ and thus the significance test of the across occasion difference in τ, yet the impact of a Λ difference depends on the size of κ, and thus on the choice of the common factor origin.

Byrne et al. (1989), who considered partial measurement invariance in multigroup designs, proposed a multistep procedure in which across group invariance of factor loadings is investigated before intercept invariance. They do not mention the constraint interaction, but in their procedure the problem is circumvented as they only constrained the intercepts of variables that have factor loadings that were found to be invariant. In effect, they use a one degree of freedom test to decide the invariance of two parameters.

In our procedure, we choose to always test the invariance of factor loadings and intercepts simultaneously, applying tests with multiple degrees of freedom. This also solves the related problem that in the presence of factor loading differences ("nonuniform bias"; Barendse et al. 2010, present AStA issue) the size of intercept differences ("uniform bias") is dependent on the (arbitrary) scale of the common factor.

2.4 Guarding against chance findings

The procedure to detect measurement bias in longitudinal data involves a very large number of possible tests, especially when there are more than two measurement occasions. The suggested steps in measurement bias detection are similar to the steps in model modification, and we should be aware of the chance capitalization problems that are associated with specification searches (MacCallum et al. 1992). Modifications should be theory driven, not data driven, and we should prevent overfitting, which would diminish the generalisability of our results.

In our procedure, we guard against chance findings in three ways. Firstly, we will only test specific hypotheses that are formulated in advance, and we will not use statistics such as the modification index and the expected parameter change (Kaplan 2000) to explore possible improvement in fit that is not associated with the hypotheses under consideration. Secondly, we will limit the number of tests by using global tests with multiple degrees of freedom, to test for the invariance of multiple parameters across all measurement occasions simultaneously. In this way, we once more forego the use of the modification index with its associated problems (Kaplan 1990). Thirdly, to prevent inflation of the family-wise error rate we will conduct all tests at Bonferroni adjusted levels of significance, in the way described by Holm (1979). To achieve this, in each step of the procedure we will test at a level of significance that is equal to the quotient of a chosen family-wise level of significance (e.g. 5%) and the number of tests under consideration.

2.5 Procedure

2.5.1 Step 1: Establishing a measurement model

Based on theory and previous research, we specify a longitudinal factor model with a pattern of factor loadings that is the same for each measurement occasion, without any constraints across measurement occasions. The matrix of residual variances and covariances consists of diagonal blocks as the residual factors of the same indicator variables are allowed to covary across measurement occasions. If this Model 1.1 does not show satisfactory fit, modification indices (or Lagrangian multiplier tests; Bollen 1989) and standardised residuals may help to guide a specification search. When modifying the model, we require equivalence of patterns of factor loadings across occasions (to consolidate equal interpretation and naming of common factors across occasions), even if this means that not all factor loadings are significant. Moreover, to guard against chance findings and the inflation of the family-wise error rate, the model should only be modified if the chi-square difference test with n_t degrees of freedom is significant at an adjusted level of significance $\alpha^* = \alpha_f/(n_z n_t)$, where α_f is the family-wise level of significance, n_z is the number of factor loadings fixed at zero for a single measurement occasion, and n_t is the number of measurement occasions. So the product of n_z and n_t is the number of tests under consideration.

Of course, to preserve a clear interpretation of the resulting model, all model modifications should have substantive justifications. We will refer to the final model in Step 1 as Model 1F.

2.5.2 Step 2: Testing measurement invariance across measurement occasions

In Step 2, the first model that we fit, Model 2.1, has the same pattern of factor loadings as Model 1F, but with across occasion equality constraints on all factor loadings and intercepts. The variances and means of the common factors are fixed for the first occasion (e.g. variances at unity and means at zero) and free for the other occasions. The matrix of residual variances and covariances has the same specification as in Step 1. The chi-square test of the difference between the fit of Model 2.1 and the fit of Model 1F may serve as a global test of the across occasion invariance of factor loadings and intercepts, but even if the test does not turn out significant we may want to consider more specific tests for each indicator variable separately.

For each of the n_i indicator variables, the fit of Model 2.1 is compared to the fit of a model in which the equality constraints on the factor loadings and intercepts associated with the indicator variable are removed. These chi-square difference tests have $(n_t - 1)(1 + n_f)$ degrees of freedom, where n_t is the number of measurement occasions and n_f is the number of free factor loadings of that indicator variable on one measurement occasion. As the number of tests is n_i, we suggest to test at an adjusted level of significance $\alpha^* = \alpha_f/n_i$. If the largest of the n_i chi-square test results is significant, then we consider the associated indicator variable as biased. The factor loadings and intercepts of the biased indicator variable remain free in Model 2.2. The fit of this model is compared to the fit of $(n_i - 1)$ other models in which the equality constraints of one of the remaining items are also cancelled. If

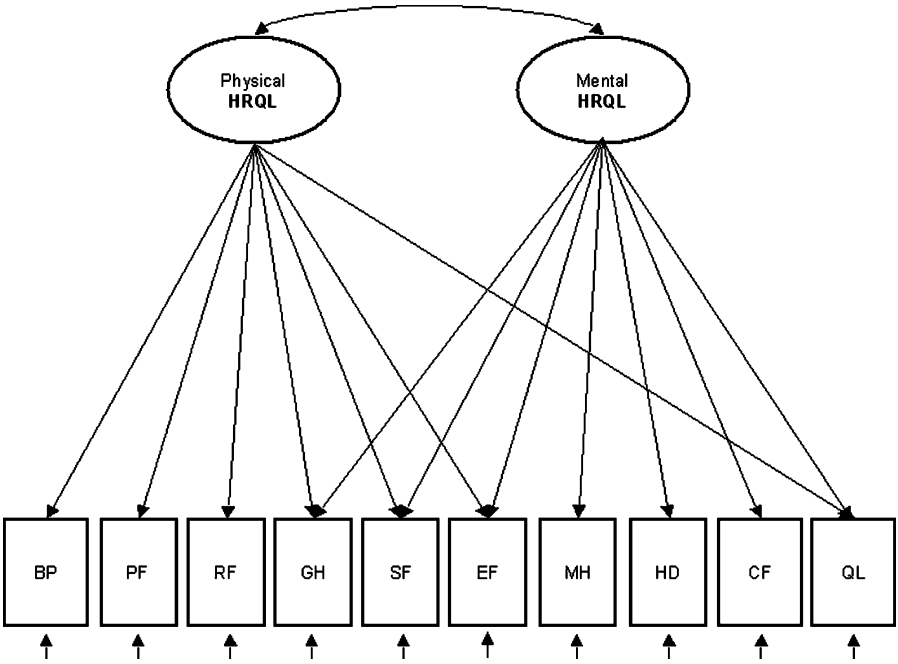

Fig. 1 Graphical display of part of Model 1F, showing the first measurement occasion variables only. Note: Only part of the model is depicted. The full model has 40 indicator variables, eight common factors, and 40 residual factors. Abbreviations: BP = bodily pain, PF = physical functioning, RF = role functioning, GH = general health perceptions, SF = social functioning, EF = energy and fatigue, MH = mental health, HD = health distress, CF = cognitive functioning, and QL = quality of life

the largest of the resulting chi-square differences turns out significant at a re-adjusted level of significance $\alpha^* = \alpha_f/(n_i - 1)$, then that indicator variable is also considered biased. This step should be repeated, re-adjusting the level of significance every time ($\alpha^* = \alpha_f/(n_i - n_2)$, where n_2 is the number of indicator variables detected as biased previously within Step 2, until no significant improvements in fit are found. Of note, if this iterative procedure leads to less than a majority of unbiased items, then this may compromise the interpretation of the findings (e.g. changes in common factor means).

After establishing the invariance or partial invariance of factor loadings and intercepts, one might also investigate the invariance of residual variances across measurement occasions, following a procedure similar to the one described above. This would be necessary if the goal is to investigate "strict" measurement invariance of the observed indicator variables (or their reliability). In the present procedure, however, the ultimate goal is to assess true change in the attribute of interest, by testing change in the common factor means. We may therefore choose to leave the residual factor variances unconstrained, as the residual factors affect neither the measurement nor the explanation of the common factors (as can be seen in Fig. 1).

The final model in Step 2 is referred to as Model 2F. If measurement bias has been detected, then one or more of the indicator variables will have varying factor loadings and intercepts across measurement occasions. Post hoc tests may aid the

interpretation of such bias, but we should be prudent when conducting such tests, in view of possible interaction constraints. Moreover, to prevent unwarranted gains in degrees of freedom, it may be better to not use the results of post hoc tests to partly re-impose constraints on the parameters concerned.

2.5.3 Step 3: Testing measurement invariance with respect to exogenous variables

Model 3.1 is obtained by extending Model 2F to include exogenous variables. All correlations between the exogenous variables and the common factors are free to be estimated and all direct effects of the exogenous variables on the observed indicator variables are fixed at zero. A non-zero effect of an exogenous variable on a particular indicator variable would indicate measurement bias in the indicator variable with respect to the exogenous variable, as the observed covariance of these two variables is not sufficiently explained by the common factor.

For each of the n_i indicator variables we test for measurement bias with respect to each of the n_e exogenous variables separately, by fitting $n_i n_e$ models in which the direct effects of the exogenous variable are set free to be estimated at all n_t measurement occasions. The associated chi-square difference tests have n_t degrees of freedom. As the number of tests is $n_i n_e$, we suggest to test at an adjusted level of significance $\alpha^* = \alpha_f / n_i n_e$. If the largest of the $n_i n_e$ chi-square test results is significant, then we consider the indicator variable as biased, leave the regressions on the exogenous variable free, and continue testing for additional measurement bias, consistently re-adjusting the level of significance ($\alpha^* = \alpha_f / n_i n_e - n_3$, where n_3 is the number of indicator variables detected as biased previously within Step 3), until no significant improvements in fit are found.

The final model in Step 3 is called Model 3F. This model can be used to assess the true change in the attributes of interest. One might evaluate such change by just inspecting the values of the common factor means, taking their standard errors or confidence intervals into consideration. If one should want to conduct significance tests, we once more suggest conducting global tests (first), at adjusted levels of significance, by fitting additional models, one for each common factor, in which the common factor means are constrained to be equal (e.g. by fixing all means at zero). The associated chi-square difference test has $n_t - 1$ degrees of freedom. If this test is conducted for each common factor separately, the adjusted level of significance is $\alpha^* = \alpha_f / n_f$, where n_f is the number of common factors per measurement occasion.

3 Illustrative example: Health survey of HIV patients

To illustrate the procedure outlined above we use data compiled from several studies that investigated the health-related quality-of-life (HRQL) of HIV/AIDS patients (Nieuwkerk 2006). The present sample comprised 403 respondents who completed an HRQL test on four semi-annual measurement occasions. The majority of the sample was male (87.5%) with a mean age of 41 ($m = 40.7$, $s = 8.7$).

The questionnaire used to assess HRQL was the well validated Medical Outcomes Study HIV Health Survey (Wu et al. 1997). The test contains 35 items that assess ten

HRQL domains: physical functioning (PF), bodily pain (BP), role functioning (RF), social functioning (SF), general health perceptions (GH), energy and fatigue (EF), health distress (HD), cognitive functioning (CF), mental health (MH) and quality of life (QL). For each domain, scale scores are calculated by summing all item scores. Item scores have been re-scaled in such a way that higher scale scores indicate better health.

Other variables included in our analysis are age, gender, time on highly active antiretroviral treatment (HAART) and the CD4-cell count, (i.e. the number of T-cells present in a cubic millimetre of blood). HAART is the treatment given to HIV and AIDS patients to postpone or slow the progression of the disease. In this case, time on HAART roughly coincides with time after diagnosis. The CD4-cell count is an indicator of the functioning of the immune system; a count less then 200 suggests progression from HIV to AIDS, and the CD4-cell count was dichotomised accordingly (Hogg et al. 2001).

The computer program LISREL was used to for maximum likelihood estimation in SEM (version 8.5, Jöreskog and Sörbom 1996). The freely available computer program NIESEM was used to calculate ECVI differences and the associated confidence intervals (Dudgeon 2003).

Table 1 includes the results of the chi-square measure of fit (CHISQ), RMSEA, and ECVI for all models discussed below, as well as CHISQ and ECVI differences for specific comparisons.

3.1 Step 1 results: Measurement model

The test manual suggests that BP, PF and RF are indicative of physical health, that MH, QL, CF and QL are indicative of mental health, and that GH, SF and EF are indicative of both physical and mental health (Wu et al. 1997). A corresponding factor model with two common factors is depicted in Fig. 1. A longitudinal version of this model with four times two common factors was fitted to the variance-covariance matrix of the four times ten HRQL scales.

The chi-square test of exact fit for this model was significant (CHISQ $= 1505.0$, $df = 640$), but the RMSEA indicated satisfactory fit (Table 1, Model 1.1). Inspection of standardised residuals and modification indices suggested cross-loadings of QL on the physical HRQL factor. Adding these parameters for each of the measurement occasions yielded a model with significantly better fit (Table 1, Model 1F). The chi-square difference test (CHISQ DIFF $= 182.0$, $df = 4$, $p < 0.0001$) is significant at the adjusted level of significance, $\alpha^* = 0.05/(7 \times 4) = 0.0018$. As the QL scale contains very general questions, we believed this to be a theoretically sound suggestion. Using this model as the new model of comparison, we checked for other significant modifications, but found none to be significant (at the re-adjusted level of significance, $\alpha^* = 0.05/(6 \times 4) = 0.0021$). As the fit of Model 1F was satisfactory (RMSEA $= 0.052$, 90% confidence interval $= [0.048, 0.056]$) and its interpretation was clear, we proceeded to Step 2.

3.2 Step 2 results: Measurement invariance across measurement occasions

In Model 2.1, all factor loadings and intercepts were constrained to be equal across the four measurement occasions. The fit of this model (Table 1, Model 2.1) was

Table 1 Overall goodness-of-fit and chi-square difference test results

Model		CHISQ (*df*)	RMSEA (90% conf. int.)	ECVI (90% conf. int.)	Comparison models	CHISQ DIFF (*df*)	*p*	ECVI DIFF (90% conf. int.)
1.1	Measurement model from manual	1505.0 (640)	0.058 (0.054; 0.062)	4.962 (4.681; 5.266)				
1F	Modified measurement model	1323.0 (636)	0.052 (0.048; 0.056)	4.532 (4.274; 4.812)	1F vs. 1.1	182.0 (4)	<0.0001	0.431 (0.325; 0.557)
2.1	Across occasion constraints on factor loadings and intercepts	1423.2 (696)	0.051 (0.047; 0.055)	4.449 (4.182; 4.738)	2.1 vs. 1F	100.2 (60)	0.0009	−0.083 (−0.143; −0.001)
2.2	Factor loadings and intercepts of HD free	1395.0 (690)	0.050 (0.046; 0.054)	4.412 (4.148; 4.697)	2.2 vs. 2.1	28.2 (6)	<0.0001	0.037 (0.004; 0.091)
2F	Factor loadings and intercepts of HD and EF free	1373.3 (681)	0.050 (0.046; 0.054)	4.408 (4.146; 4.691)	2F vs. 2.2	21.7 (9)	0.0099	0.004 (−0.021; 0.051)
2.4	Factor loadings and intercepts of HD, EF, and MH free	1365.8 (675)	0.050 (0.047; 0.054)	4.422 (4.161; 4.705)	2.4 vs. 2F	7.5 (6)	0.2771	−0.015 (−0.016; 0.015)
2.5	As Model 2F but with constrained residual variances	1478.0 (711)	0.052 (0.048; 0.056)	4.502 (4.229; 4.797)	2.5 vs. 2F	104.7 (30)	<0.0001	0.094 (0.023; 0.186)
3.1	As Model 2F but with addition of exogenous variables	1563.2 (809)	0.047 (0.044; 0.051)	5.149 (4.871; 5.449)				
3.2	Free direct effects of CD4-cell count on EF	1534.8 (805)	0.047 (0.044; 0.050)	5.101 (4.826; 5.398)	3.2 vs. 3.1	28.4 (4)	<0.0001	0.048 (0.014; 0.104)
3F	Free direct effects of CD4-cell count on EF and RF	1516.3 (801)	0.047 (0.043; 0.051)	5.077 (4.805; 5.372)	3F vs. 3.2	18.5 (4)	0.0010	0.024 (−0.001; 0.070)
3.4	Free direct effects of CD4-cell count on EF and RF, and of Gender on HD	1506.0 (797)	0.046 (0.043; 0.050)	5.074 (4.803; 5.367)	3.4 vs. 3F	10.3 (4)	0.0357	0.003 (−0.011; 0.039)

Note: $N = 403$

significantly worse when compared to Model 1F (CHISQ DIFF = 100.2, $df = 60$, $p = 0.0009$). In the series of models that was fitted next, the model in which the equality constraints were removed for HD showed the best fit (Table 1, Model 2.2). The chi-square difference test (CHISQ DIFF = 28.2, $df = 6$, $p < 0.0001$) turned out significant at the adjusted level of significance, $\alpha^* = 0.05/10 = 0.005$. Retaining the additional parameters, Model 2.2 was used as the comparison model in the evaluation of a new series of models. The model with the parameters of EF freed yielded the largest chi-square difference (CHISQ DIFF = 21.7, $df = 9$, $p = 0.0099$). As the CHISQ DIFF was significant at the re-adjusted level of significance, $\alpha^* = 0.05/9 = 0.0056$, we retained the additional parameters and used this model for subsequent comparisons. However, none of the models in the next series showed a significant improvement of fit (the largest improvement was found for a model with free MH parameters, CHISQ DIFF = 7.5, $df = 6$, $p = 0.2771$).

With HD (health distress) we see that the intercept for the first measurement occasion is notably lower than the other intercepts. Apparently, it is more difficult for respondents to answer positively to the HD items (to show less health distress and score high on the HD scale) when they have just entered the research and complete the HRQL test for the first time. In subsequent administrations of the HRQL test, respondents score higher on the HD scale, relative to their Mental HRQL. The factor loadings of HD on Mental HRQL go up and down with time, which makes the bias difficult to interpret.

With EF (energy and fatigue), the intercepts seem to decrease, which would indicate that with time, it becomes more difficult to agree with the EF items (i.e. to do well and score high on the EF scale), relative to the respondent's Physical HRQL and Mental HRQL. That is, when general HRQL improves, EF does not improve as much. The factor loadings of EF on Physical HRQL and Mental HRQL show a pattern that suggests that with time, the answers to the energy and fatigue (EF) items become less indicative of the respondents' Physical HRQL and more indicative of the respondents' Mental HRQL, but these differences are very small.

Model 2F was also used as the comparison model in a test of invariance of the residual variances. This test turned out highly significant (CHISQ DIFF = 104.7, $df = 30$, $p < 0.0001$, Table 1). As invariance of residual variances is not required for valid assessment of change in the common factor means, we did not follow up with tests of partial invariance of the residual variances.

3.3 Step 3 results: Measurement invariance with respect to exogenous variables

We included variables that are supposed to be related to HRQL of HIV/AIDS patients and that may induce bias in the test scores. In Model 3.1, age, gender, CD4-cell count and time on HAART were correlated with the common factors Physical and Mental HRQL at all four measurement occasions, but the four new variables were not allowed to directly affect the observed indicator variables of Physical and Mental HRQL. Part of Model 3.1 is depicted in Fig. 2, only showing the HRQL variables of the first measurement occasion. The fit of Model 3.1 was satisfactory (CHISQ = 1535.8, $df = 809$, RMSEA = 0.047, Table 1).

We used Model 3.1 as the comparison model in the first iteration of tests for bias with respect to the exogenous variables. The largest chi-square difference was found

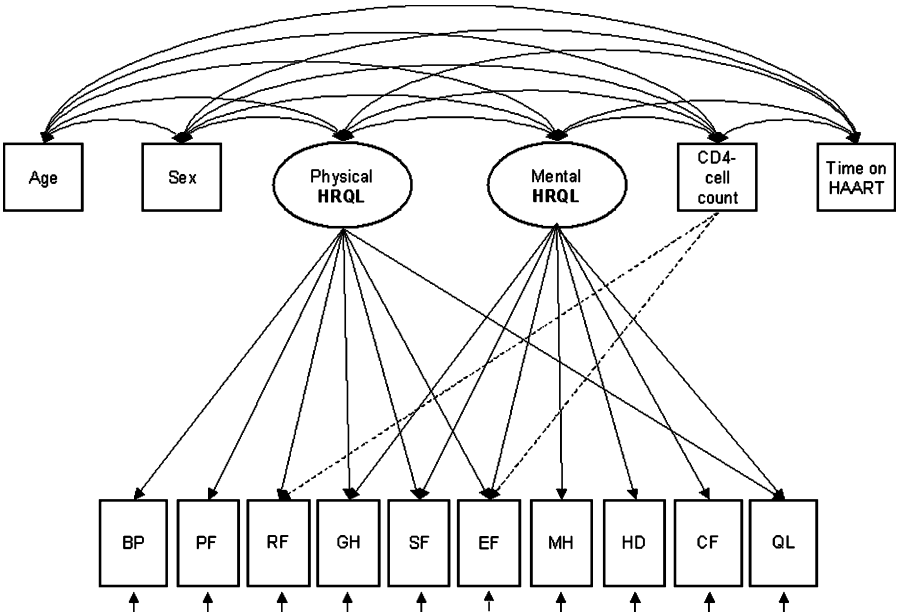

Fig. 2 Graphical display of part of Model 3.1, showing the first occasion variables only. Note: Only part of the model is depicted. The full model has 40 indicator variables, eight common factors, 40 residual factors, and four exogenous variables. The *dashed arrow* represents measurement bias in RF and EF with respect to CD4-cell count

for the direct effects of CD4-cell count on the EF indicators (CHISQ DIFF = 28.4, $df = 4$, $p < 0.0001$), and it was significant at the adjusted level of significance of $\alpha^* = 0.05/(10 \times 4) = 0.0013$. We therefore retained these direct effects in the next comparison model, Model 3.2, to test for additional direct effects of exogenous variables on observed indicators. The largest chi-square difference was found for the direct effects of CD4-cell count on the RF indicators (CHISQ DIFF = 18.5, $df = 4$, $p = 0.0010$), which was only barely significant at the re-adjusted level of significance of $\alpha^* = 0.05/39 = 0.0013$. In the next iteration, none of the resulting chi-square difference tests turned out significant. The largest chi-square difference was associated with direct gender effects on HD (CHISQ DIFF = 10.3, $df = 4$, $p = 0.0357$).

In order to interpret apparent measurement bias, we have to take the relationships of the exogenous variables with the common factors into account as well. The correlations from Model 3F are given in Table 3. To check whether apparent measurement bias is consistent over time, we compared the fit of Model 3F with a model in which the direct effects of CD4-cell count on EF were constrained to be equal. The resulting model fitted almost as well as Model 3F (Table 1, CHISQ DIFF = 1.4, $df = 3$, $p = 0.7055$). In this model, the direct effect of CD4-cell count on EF was estimated at -0.39 ($se = 0.09$) for all four measurement occasions, indicating that a low CD4-cell count (indicative of having AIDS rather than HIV) affects the respondent's energy and fatigue (lower EF scores) in another way than would be expected on account of the positive correlation between CD4-cell count and Physical HRQL at the first mea-

Table 2 Intercepts and factor loadings (selected parameter estimates from Model 2F)

HRQL scale	Intercepts for four measurement occasions	Factor loadings Physical HRQL for four measurement occasions	Factor loadings Mental HRQL for four measurement occasions
BP	7.83 (0.10)	1.64 (0.08)	0
PF	7.86 (0.10)	1.64 (0.08)	0
RF	6.60 (0.18)	2.86 (0.14)	0
GH	5.02 (0.10)	1.24 (0.08)	0.58 (0.07)
SF	7.51 (0.11)	1.68 (0.09)	0.35 (0.07)
EF	6.20/6.17/6.06/6.01 (0.10/0.10/0.10/0.10)	1.09/0.86/0.85/0.89 (0.08/0.08/0.08/0.07)	0.90/1.10/1.03/0.97 (0.08/0.09/0.08/0.07)
MH	7.00 (0.09)	0	1.62 (0.07)
HD	7.54/7.76/7.68/7.74 (0.10/0.10/0.10/0.10)	0	1.32/1.58/1.38/1.50 (0.08/0.09/0.08/0.08)
CF	7.60 (0.09)	0	1.03 (0.06)
QL	6.80 (0.10)	1.07 (0.07)	1.10 (0.07)

Notes: standard errors are given within parentheses; a single entry indicates that the parameter estimate is constrained to be equal across the four measurement occasions; to save space, common factor means, variances, and covariances, and residual variances and covariances are not shown

surement occasions (Table 3). In other words, a CD4-cell count indicative of AIDS is associated with worse Physical HRQL, but not as much with EF.

We also tested whether the measurement bias in RF (role functioning) with respect to CD4-cell count was consistent over time, but this appeared not to be the case (CHISQ DIFF $= 17.5$, $df = 3$, $p = 0.0006$). The direct effects of CD-4 cell count on RF varied across measurement occasions, being negative at the first occasion and positive on the other occasions, -0.60 ($se = 0.37$), 0.92 ($se = 0.34$), 0.39 ($se = 0.35$), and 0.08 ($se = 0.31$). In spite of testing at an adjusted level of significance, this may be a chance result. In fact, due to the large standard errors, only the second occasion 0.92 effect is significant at the 0.05 level.

With measurement bias accounted for, Model 3F was used to evaluate change in Physical and Mental HRQL. The common factor means and their standard errors are given in Table 3. Both Physical HRQL and Mental HRQL improve after the first

Table 3 Common factor means and correlations (selected parameter estimates from Model 3F)

	Physical HRQL T1	Mental HRQL T1	Physical HRQL T2	Mental HRQL T2	Physical HRQL T3	Mental HRQL T3	Physical HRQL T4	Mental HRQL T4
Common factor means								
Means	0	0	0.16	0.06	0.15	0.18	0.16	0.14
Standard errors			(0.04)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
Common factor correlations								
Physical HRQL T1								
Mental HRQL T1	0.43	1.00						
Physical HRQL T2	0.73	0.32	1.00					
Mental HRQL T2	0.38	0.71	0.59	1.00				
Physical HRQL T3	0.72	0.28	0.83	0.44	1.00			
Mental HRQL T3	0.39	0.72	0.49	0.81	0.57	1.00		
Physical HRQL T4	0.60	0.26	0.72	0.37	0.80	0.41	1.00	
Mental HRQL T4	0.34	0.65	0.43	0.67	0.45	0.76	0.60	1.00
Gender	0.02	0.04	−0.07	0.02	−0.05	0.04	0.02	0.10
Age	−0.07	0.11	−0.08	0.03	−0.11	0.01	−0.17	−0.03
CD4-cell count	0.31	0.07	0.11	−0.04	0.12	−0.01	0.03	−0.02
Time on HAART	−0.08	0.18	−0.25	0.01	−0.21	−0.01	−0.28	−0.07
	Gender	Age	CD4-cell count	Time on HAART				
Correlations between exogenous variables								
Gender	1.00							
Age	−0.19	1.00						
CD4-cell count	0.04	0.00	1.00					
Time on HAART	−0.02	0.17	0.19	1.00				

measurement occasion. Perhaps patients get used to the idea of being HIV infected or even of having AIDS, and learn to live with it, so that the disease does not affect their HRQL as much when they completed the HRQL test for the first time. The correlations between the exogenous variables and HRQL are generally small, except for the correlation between CD4-cell count and the first occasion measurement of Physical HRQL (0.30, if you do not have AIDS yet then you are doing better physically) and the consistent negative correlations between time on HAART and Physical HRQL. The longer the respondents receive HAART, the worse their Physical HRQL, which is understandable in view of the invasive side effects of HAART (Nieuwkerk 2006).

4 Discussion

In the application of the measurement bias detection procedure to the longitudinal HRQL data of HIV/AIDS patients, we found four examples of measurement bias. First, we found the factor loadings and intercepts of HD (health distress) and EF (energy and fatigue) not to be invariant across measurement occasions and, second, we found direct effects of CD4-cell count on EF and RF (role-functioning). The first two findings of measurement bias are considered as response shift by definition, as the measurement invariance is violated by the time of the measurement occasion. However, upon inspecting the HD and EF parameter estimates (Table 2) there did not appear to be an obvious substantive explanation for the changes in the factor loadings of HD. The other two findings of measurement bias are considered as response shift only if they vary with time. The bias in EF with respect to CD4-cell count is consistent over time and therefore not considered as response shift. The bias in RF with respect to CD4-cell count did vary with time, but again, it was difficult to provide a substantive explanation for this so-called response shift. Perhaps some of our results are chance findings, despite our best attempts to guard against such findings.

The Bonferroni adjustment of the level of significance guards against inflation of the family-wise error rate, but the chi-square difference test can still be affected by model complexity and sample size. In a simulation study, Cheung and Rensvold (2002) considered various alternatives to the chi-square difference test for testing across group constraints in multi-group factor analysis, and recommended inspection of differences in Bentler's (1990) comparative fit index (among others). In our longitudinal factor analysis, we complemented the chi-square differences with ECVI differences, really only in order to provide additional information about the necessity of further modifications that cannot be substantively justified. In the present analyses, the ECVI differences generally agreed with the chi-square difference tests at Bonferroni adjusted levels of significance. One notable exception was that according to the 90% confidence interval of the ECVI difference, the fit of Models 2.2 and 2F was essentially equivalent, suggesting that constraints on EF factor loadings and intercepts could have been retained.

It should be noted that most response shift researchers in substantive areas of psychology contend that response shifts are the result of some catalyst event, such as an intervention in educational research (Howard et al. 1979), or a health state change in medical research (Sprangers and Schwartz 1999). In the HRQL study of HIV/AIDS

patients, there is not a well defined event that all respondents have in common, other than having been diagnosed with HIV or AIDS some time ago. However, the time since diagnosis and the time on HAART vary greatly across patients and cannot be considered true catalysts. The one thing all patients have in common is that they participate in the HRQL study, and that they complete HRQL tests every half year. The test taking itself can have an effect on their response behaviour, which may change with time. The patients may become more accustomed to both their disease and taking the test, which perhaps induces a response shift. It should also be noted that most work on response shift in substantive psychological research was not aimed at investigating measurement invariance, but rather at explaining paradoxical intervention effects. Seeing that research into response shift was hampered by researchers having different conceptions of response shift, Oort (2005b) proposed to formally define response shift as a special case of measurement bias, although some researchers may still have another perspective on response shift (Oort et al. 2009).

As is illustrated by the empirical example, Step 2 and Step 3 of the detection procedure are laborious and time consuming. Especially if the numbers of observed variables and exogenous variables are large, these two steps involve the fitting of numerous models, in order to evaluate the chi-difference tests. An advantage of using modification indices is that, within each iteration, the researcher only has to fit a single model. Therefore, although perhaps less sound (Kaplan 1990), we explored the use of the modification index as an alternative to the global tests with multiple degrees of freedom.

When we evaluated the modification indices with the Bonferroni adjusted levels of significance, none of the findings were significant because of the large number of tests under consideration (e.g. 120 in Step 2). When testing at less conservative levels of significance, for example by considering tests of intercept constraints first and factor loading constraints second, or by simply raising the family-wise level of significance, there was a number of modification indices that reached significance. However, as multiple modification indices were about equally large, the choice of which constraint to remove first seemed arbitrary, yet highly consequential for the removal of constraints in subsequent iterations, leading to very different conclusions. In addition, we also had to be careful not to run into constraint interactions. Still, the most important problem with relying on modification indices and less conservative testing was that many of the modifications were difficult to interpret and that the number of iterations grew very large. Saris et al. (2009) suggest only modifying models if the modification indices are associated either with moderate (instead of high) statistical power or with substantial expected parameter changes. When statistical power is high, one can only rely on substantive arguments for modification (ibidem), which we did, as in the present analyses the power to find medium sized differences was consistently above 99%.

In such situations, the decision making becomes increasingly subjective, as researchers will have to base their decisions between modifications and when to stop modifications on the interpretability of the different modifications. It is therefore highly likely that different researchers, with different substantive knowledge and different interpretation skills, will end up with different conclusions when analysing the same data. As can be seen from the procedure using modification indices, subjectivity in measurement bias detection influences whether and where bias is found. Not

all researchers may want to test every possible combination of tenable equality constraints. When this is the case, a priori hypotheses driven by theory should be stated before analysis and only these tests should be conducted. Under these circumstances, chance findings may further be reduced and more generalisable results found.

The problems associated with devising an objective procedure for measurement bias detection is common to specification searches in general. Bollen (2000): "Modelling strategies are subject to debate for virtually all statistical procedures. Witness the sharp disagreements over stepwise regression, the interpretation of clusters in cluster analysis, or the identification of outliers and influential points. The largely objective basis of statistical algorithms does not remove the need for human judgment in their implementation." Similarly, when investigating measurement invariance, it is impossible to completely remove the element of human judgement. This is certainly true for the substantive interpretation of apparent measurement bias. However, we think that the procedure presented in this paper, with its safeguards against chance findings, at least helps to more objectively decide which measurements are biased and which are not.

Acknowledgement The authors thank P.T. Nieuwkerk (Medical Psychology, Academic Medical Centre, University of Amsterdam) for making the quality-of-life data available for secondary analysis.

References

Barendse, M.T., Oort, F.J., Garst, G.J.A.: Using restricted factor analysis with latent moderated structures to detect uniform and non-uniform measurement bias; A simulation study. Adv. Stat. Anal. (2010). doi:10.1007/s10182-010-0126-1

Bentler, P.M.: Comparative fit indexes in structural models. Psychol. Bull. **107**, 238–246 (1990)

Bollen, K.A.: Structural Equations with Latent Variables. Wiley, New York (1989)

Bollen, K.A.: Modeling strategies: in search of the Holy Grail. Struct. Equ. Modeling **7**(1), 74–81 (2000)

Browne, M.W., Cudeck, R.: Alternative ways of assessing model fit. Sociol. Methods Res. **21**, 230–258 (1992)

Byrne, B.M., Shavelson, R.J., Muthén, B.: Testing for the equivalence of factor covariance and mean structures—the issue of partial measurement invariance. Psychol. Bull. **105**, 456–466 (1989)

Cheung, G.W., Rensvold, R.B.: Evaluating goodness-of-fit indices for testing measurement invariance. Struct. Equ. Modeling **9**, 233–255 (2002)

Dudgeon, P.: NIESEM: A computer program for calculating noncentral interval estimates (and power analysis) for structural equation modeling [Computer software] (2003)

Golembiewski, R.T., Billingsley, K., Yeager, S.: Measuring change and persistence in human affairs—types of change generated by OD designs. J. Appl. Behav. Sci. **12**, 133–157 (1976)

Hogg, R.S., Yip, B., Chan, K.J., et al.: Rates of disease progression by baseline CD4 cell count and viral load after initiating triple-drug therapy. J. Am. Med. Assoc. **286**, 2568–2577 (2001)

Holm, S.: A simple sequentially rejective multiple test procedure. Scand. J. Statist. **6**, 65–70 (1979)

Howard, G.S., Ralph, K.M., Gulanick, N.A., et al.: Internal invalidity in pretest-posttest self-report evaluations and a re-evaluation of retrospective pretests. Appl. Psychol. Meas. **3**, 1–23 (1979)

Jöreskog, K.G., Sörbom, D.: LISREL 8 User's Guide, 2nd edn. Scientific Software International, Inc., Chicago (1996)

Kaplan, D.: Evaluating and modifying covariance structure models: a review and recommendation. Multivar. Behav. Res. **25**, 137–155 (1990)

Kaplan, D.: Structural Equation Modeling. Sage, Thousand Oaks (2000)

MacCallum, R.C., Roznowski, M., Necowitz, L.B.: Model modifications in covariance structure analysis: the problem of capitalization on chance. Psychol. Bull. **111**, 490–504 (1992)
Mellenbergh, G.J.: Item bias and item response theory. Int. J. Educ. Res. **13**, 127–143 (1989)
Meredith, W.: Measurement invariance, factor-analysis and factorial invariance. Psychometrika **58**, 525–543 (1993)
Meredith, W., Horn, J.: The role of factorial invariance in modeling growth and change. New methods for the analysis of change. In: Collins, L.M., Sayer, A.G. (eds.) New Methods for the Analysis of Change. Decade of Behavior, pp. 179–200. APA, Washington (2001)
Nieuwkerk, P.T.: Highly active antiretroviral therapy for HIV-1 infection: Patients' quality of life and treatment adherence. Dissertation, Amsterdam: University of Amsterdam (2006)
Oort, F.J.: Theory of violators: assessing unidimensionality of psychological measures. In: Steyer, R., Wender, K.F., Widaman, K.F. (eds.) Psychometric Methodology, pp. 377–381. Fischer, Stuttgart (1991)
Oort, F.J.: Using structural equation modeling to detect response shifts and true change. Qual. Life Res. **14**, 587–598 (2005a)
Oort, F.J.: Towards a formal definition of response shift (in reply to G.W. Donaldson). Qual. Life Res. **14**, 2353–2355 (2005b)
Oort, F.J., Visser, M.R.M., Sprangers, M.A.G.: Formal definitions of measurement bias and explanation bias clarify measurement and conceptual perspectives on response shift. J. Clin. Epidemiol. **62**, 1126–1137 (2009)
Saris, W.E., Satorra, A., van der Veld, W.M.: Testing structural equation models or detection of misspecifications?. Struct. Equ. Modeling **16**, 561–582 (2009)
Sayer, A.G., Cumsille, P.E.: Second-order latent growth models. New methods for the analysis of change. In: Collins, L.M., Sayer, A.G. (eds.) New Methods for the Analysis of Change. Decade of Behavior, pp. 179–200. APA, Washington (2001)
Schmitt, N., Kuljanin, G.: Measurement invariance: review practice and implications. Hum. Resour. Manage R **18**, 210–222 (2008)
Sprangers, M.A.G., Schwartz, C.E.: Integrating response shift into health-related quality of life research: a theoretical model. Soc. Sci. Med. **48**(11), 1507–1515 (1999)
Steiger, J.H., Shapiro, A., Browne, M.W.: On the multivariate asymptotic distribution of sequential Chi-square statistics. Psychometrika **50**(3), 253–263 (1985)
Vandenberg, R.J., Lance, C.E.: A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. Organ. Res. Methods **3**, 4–70 (2000)
Wu, A.W., Revicki, D.A., Jacobson, D., Malitz, F.E.: Evidence for reliability, validity and usefulness of the Medical Outcomes Study HIV Health Survey (MOS-HIV). Qual. Life Res. **6**, 481–493 (1997)