Qual Life Res (2011) 20:1527–1540
DOI 10.1007/s11136-010-9844-9

Using structural equation modeling to detect response shift in performance and health-related quality of life scores of multiple sclerosis patients

Bellinda L. King-Kallimanis · Frans. J. Oort · Sandra Nolte · Carolyn E. Schwartz · Mirjam A. G. Sprangers

Accepted: 28 December 2010 / Published online: 19 January 2011

Abstract

Purpose To illustrate how structural equation modeling (SEM) can be used for response shift detection with random measurement occasions and health state operationalized as fixed group membership (Study 1) or with fixed measurement occasions and health state operationalized as time-varying covariates (Study 2).

Methods In Study 1, we explored seven items of the Performance Scales measuring physical and mental aspects of perceived disability of 771 stable, 629 progressive, and 1,552 relapsing MS patients. Time lags between the three measurements varied and were accounted for by introducing time since diagnosis as an exogenous variable. In Study 2, we considered the SF-12 scales measuring physical and mental components of HRQoL of 1,767 patients. Health state was accounted for by exogenous variables relapse (yes/no) and symptoms (worse/same/better).

Results In Study 1, progressive and relapsing patients reported greater disability than stable patients but little longitudinal change. Some response shift was found with stable and relapsing patients. In Study 2, relapse and symptoms were associated with HRQoL, but no change and only little response shift was found.

Conclusions While small response shifts were found, they had little impact on the evaluation of true change in performance and HRQoL.

Keywords Response shift · Structural equation modeling · Health-related quality of life · Multiple sclerosis patients · Measurement bias

Introduction

Measurement in health research relies heavily on self-report data. Self-report data collected in longitudinal studies are often difficult to interpret due to respondents' changing standards, values, or conceptualization of the target construct. This phenomenon is referred to as 'response shift'. We distinguish three types of response shift: (1) *recalibration* of respondents' internal standards of measurement, (2) *reprioritization* of respondents' values, and 3) *reconceptualization* of the target construct [1]. Each of these types of response shift can be operationalized within structural equation modeling (SEM) [2, 3].

Several operationalizations of response shift have been proposed. Generally, response shift can be defined either as bias in the measurement of the attribute of interest or as bias in the explanation of the attribute [4]. In this paper, we will focus only on response shift in measurement. From this perspective, bias is not considered as noise but rather as systematic differences in patients' scores that are not

B. L. King-Kallimanis · Frans. J. Oort · M. A. G. Sprangers
Department of Medical Psychology, Academic Medical Centre,
University of Amsterdam, Amsterdam, The Netherlands

Frans. J. Oort (✉)
Department of Education, Faculty of Social and Behavioural
Sciences, University of Amsterdam, Nieuwe Prinsengracht 130,
1018 VZ Amsterdam, The Netherlands
e-mail: f.j.oort@uva.nl

S. Nolte
Association of Dermatological Prevention, Hamburg,
Germany

C. E. Schwartz
DeltaQuest Foundation, Concord, MA, USA

C. E. Schwartz
Departments of Medicine and Orthopaedic Surgery,
Tufts University School of Medicine, Boston, MA, USA

fully explained by true differences in the attribute of interest (e.g., health-related quality-of-life (HRQoL)), but also by differences in other variables (e.g., other patient expectations, adaptation). Response shift is considered as a measurement bias that changes with time of measurement in longitudinal research (see Oort et al. [2, 4, 5]).

With SEM, we can detect such measurement bias with respect to time of measurement in longitudinal designs, group membership in multigroup designs, or any other exogenous variable. For example, the effects of health state on the course of HRQoL can be modeled by dividing the sample into healthy and non-healthy subgroups, or by including an indicator of health state as an exogenous variable. If exogenous variables are included in a longitudinal model, they can be static (e.g., diagnosis) or they can vary across measurement occasions (e.g., depression scale scores). Additionally, different longitudinal structures (e.g., growth, autoregression) can be investigated with latent variables.

As explained in a companion paper by Schwartz et al. [6], this paper is one in a series investigating response shift in multiple sclerosis (MS) patients using different methods. Here, we demonstrate how SEM can be used to detect response shift. We aim to illustrate the flexibility of SEM by investigating response shift in two studies. In Study 1, we investigate the performance disabilities in MS patients by taking the first three measurement occasions with varying time lags across patients. We investigate health status by distinguishing between three pre-defined and known groups of MS patients (i.e. stable, progressive, and relapsing) and use these groups in a multigroup analysis. In Study 2, we investigate HRQoL in MS patients by selecting measurement occasions with fixed time lags. Health status is taken into account by introducing time-varying health status indicators as exogenous variables. In both studies, we will investigate change and response shift with respect to health status.

Method

Data

Analyses in this paper utilize data from the North American Research Committee on Multiple Sclerosis (NARCOMS) project registry. The NARCOMS registry was established in 1993 to collect biannual data on MS patients' status. The main aim of the registry is to make these data available for the wider community, in particular researchers, to increase knowledge about MS.

Study 1

Response shift in performance disability is investigated using the intake questionnaire and the two subsequent follow-up questionnaires. As the timing of the measurement occasions varies across patients, they are considered random. On average, the first two measurement occasions are 1.04 (SD = 0.79) years apart, and the second and third measurement occasions are 0.88 (SD = 0.73) years apart.

Study 2

In Study 2, three other measurement occasions are used. Since the intake survey does not include the HRQoL questionnaire, we took the first three measurement occasions that included the HRQoL questionnaire and that were evenly spaced in time (about 6 months apart). These occasions are considered as fixed. On average, the first measurement occasion in this study is 3.07 (SD = 1.97) years from intake.

Variables

In the NARCOMS registry, a number of demographic, clinical, and psycho-social measures are collected. In Study 1, we investigate change and response shift in 'performance disability' as measured by the Performance Scales [7]. In Study 2, we investigate change and response shift in 'HRQoL' as measured by the SF-12 [8]. In both studies, we include demographic variables (age and sex) and clinical variables (time since diagnosis and health state). These variables are used to investigate additional measurement bias in the observed variables of the Performance Scales and the SF-12.

Performance Disability

The Performance Scales [7] originally included eight items. As the visual disability item was not consistently included as part of the NARCOMS survey, we use seven items of disability. Items were scored on a 6-point (or 7-point in case of mobility) Likert scale (0 = Normal; 5 = Totally disabled) and represent performance disability with respect to mobility, hand function, fatigue, cognition, bladder/bowel, sensory, and spasticity. Higher scores indicate greater disability. When less than three item responses were missing, values were imputed using the expected maximization (EM) algorithm [9].

HRQoL

The SF-12 [8] was used to measure two components of HRQoL: Mental (MENT) and Physical (PHYS). Eight scales are created from the 12 items including physical functioning (PF), role limitations because of physical health (RP), bodily pain (BP), general health (GH), vitality (VT), social functioning (SF), role limitations because of emotional problems (RE), and mental health (MH). Higher scores indicate better HRQoL. The scales—not the items—are the focus of our analysis. When less then five subscale values were missing, values were imputed using the EM algorithm [9].

Health State

In Study 1, three groups of patients with different health states were created based on their answers to relapse and symptom change questions at baseline and follow-up measurements. The three groups are defined as follows: 'stable', patients with no relapses and symptoms that remained unchanged or improved; 'progressive', patients who relapsed and whose symptoms continued to get worse; 'relapsing', those who experienced a relapse but whose symptoms remained unchanged or improved. In Study 2, two items were used to measure health state: 'relapse in the past 6 months' (1 = yes, 0 = no/unsure) and 'symptoms compared to 6 months ago' (1 = much worse; 7 = much better). Both items were administered at each measurement occasion and can thus be included as time-varying covariates.

Other Variables

'Age', 'sex', 'newly diagnosed', and 'time since diagnosis' are also included in the analyses. At the first measurement occasion of Study 1, we distinguish between patients who are newly diagnosed (diagnosis the same year as joining the NARCOMS registry) and patients whose diagnosis year was different from the year of joining NARCOMS or their diagnosis year is unknown. In Study 2, 'time since diagnosis' is treated as a continuous variable that is calculated as the difference between the first measurement occasion and the year of diagnosis. Patients with an unknown year of diagnosis are excluded.

Study 1 analysis

In both studies, the analysis has three steps: Establishing a measurement model (Step 1), testing invariance of model parameters across measurement occasions (Step 2), and testing invariance with respect to exogenous variables (Step 3). Each step is outlined below, with similarities and differences between the two studies highlighted. All analyses were carried out with LISREL 8.54. See Appendix 1 for a more detailed description of the methods; syntax files are available upon request.

Step 1: Establishing a measurement model

The Performance Scales were originally reported to measure a unidimensional construct [7]. If the corresponding confirmatory factor analysis (CFA) model does not fit, then exploratory factor analysis is used to determine an alternative model, before continuing with CFA. Maximum likelihood estimation is used for parameter estimation. We assess the overall fit of our models with the chi-square test of exact fit, the root mean square error of approximation (RMSEA) [12], the expected cross-validation index (ECVI) [12], the comparative fit index (CFI) [10], and the Tucker Lewis index (TLI) [11]. A non-significant chi-square value indicates good fit. However, as it is sensitive to small deviations between model and data, especially when sample size is large, we also consider approximate fit indices. RMSEA < 0.08 indicates satisfactory fit; RMSEA < 0.05 indicates close fit. ECVI cannot be used as a stand-alone index but can be used to compare alternative models; a smaller ECVI value indicates better model fit [12]. Finally, both the CFI and the TLI assess the improvement in fit from a null model that assumes no relationships between variables. Values of $> .90$ for the TLI and values $> .95$ for the CFI indicate reasonable fit of the model to data. If a new model is specified, the change in model fit is assessed with the chi-square difference test and the ECVI difference test [12].

Step 2: Testing invariance across measurement occasions

In this step, we take the final model of Step 1 and simultaneously constrain all factor loadings and intercepts to be equal across measurement occasions and groups. Across-occasion invariance (no measurement bias) of factor loadings and intercepts is assessed by comparing this model with the final model of Step 1 using the chi-square difference test. A significant result provides evidence for response shift. However, if the test result is not significant, we still investigate response shift, as a single, yet substantially important response shift may not cause significant deterioration in the overall model fit.

To detect measurement bias, we examine modification indices and the standardized expected parameter changes (SEPC) [13]. If both are large, we expect significant improvement in the overall model and substantial change in the parameter estimate(s). As there are a large number of modification indices to consider, we stop investigating modification indices when none are greater than a Bonferroni-adjusted critical value [14] of 12.83. For SEPC, we consider > 0.10 significant [15]. The effect size [16] of possible response shifts will be evaluated in comparison with Cohen's *d* effect sizes of observed and true change [2].

Step 3: Testing invariance with respect to exogenous variables

The first model in this step includes 'age', 'sex', 'newly diagnosed', 'time between measurement occasions 1 and 2', and ‘time between measurement occasions 2 and 3’ as additional exogenous variables. These five exogenous variables correlate with each other and with the common factors, but their relationship with the observed items should be fully explained through these correlations. If large modification indices and SEPCs are present, this indicates the presence of bias. In case of direct effects changing over time, we consider this measurement bias as response shift.

Study 2 analysis

In Study 2, we took the same steps as in Study 1. All analyses were carried out with Mx [17]. See Appendix 2 for a detailed description of the methods; syntax files are available upon request.

Step 1: Establishing a measurement model

The first goal is to find a satisfactory measurement model for the SF-12. We begin with the measurement model comprising two common factors: PHYS and MENT HRQoL. If this model does not fit, we use modification indices and standardized residuals [13, 18] to identify misspecification and to develop an alternative model. As in Study 1, possible model modifications are assessed using the chi-square difference test and ECVI difference test. Overall model fit is assessed using the same statistics as used in Study 1.

Step 2: Testing invariance across measurement occasions

All factor loadings and intercepts of the final model of Step 1 are simultaneously constrained to be equal across measurement occasions, like in Study 1. To detect response shift, we use a different search strategy from Study 1 where we test individual constraints. Here, we follow the procedure outlined in King-Kallimanis et al. [19], relying on a smaller number of global tests that free multiple constraints simultaneously. In this study, we use eight global tests, one for each observed scale. The fit of each of these eight new models is compared to the fully constrained model using the chi-square difference test (at adjusted significance level) [14] and scaled observed parameter changes (OPC). After running the eight tests, the observed scale producing the largest OPC in combination with a significant chi-square difference test is interpreted as response shift. We continue iteratively, retesting the remaining scales, until no large OPC with a significant chi-square difference test is found. Corresponding to Cohen’s small effect sizes, we consider an OPC indicating a standardized difference of 0.1 between factor loadings or 0.2 between intercepts to be large [16].

Step 3: Testing invariance with respect to exogenous variables

We extend the final model of Step 2 to include ‘age’, ‘sex’, ‘time since diagnosis’, ‘relapse in the past 6 months’, and ‘symptom change in the last 6 months’ as exogenous variables. To test for response shift, we fit new models where we include direct effects of the exogenous variables on the observed scales. The impact of these direct effects is assessed with OPCs and the chi-square difference test. If the largest effects are significant, we leave these parameters free to be estimated and repeat the process until no significant improvements are found. Once any biases have been accounted for, this final model can be used to assess true change in the attribute of interest using the same formula used in Study 1 [2].

Results

Study 1 results

In the analysis of the Performance Scales items, we distinguished between 771 stable patients (26.1%), 629 progressive patients (21.3%), and 1,552 relapsing patients (52.6%). See Table 1 for sample characteristics and Table 2 for the Performance Scales item means.

Step 1: Establishing a measurement model

The reported unidimensional structure of the Performance Scales yielded satisfactory fit (Table 3, Model 1.1.1). However, model fit could be improved upon. Exploratory factor analysis suggested a two-dimensional model, and in CFA this model yielded substantially better fit for the Performance Scales than a one-dimensional model (Table 3, 1.1.2). The two dimensions were interpreted as (1) Visible Disability (mobility, spasticity, bladder), describing the most visible and stigmatizing symptoms that may make one home bound; and (2) Internal Disability (hand function, fatigue, sensory, cognition), relating to an internal, more subjective experience.

Step 2: Testing measurement invariance across measurement occasions

The equality constraints imposed in this step led to a significant deterioration in fit, suggesting the presence of response shift (Table 3, Model 1.2.1). The modification indices and SEPCs suggested first removing the equality constraint on the intercept of ‘sensory’ for the stable group at the first measurement occasion (χ^2diff(1) = 56.8, $P < 0.001$) and successively the intercept of ‘sensory’ for

Table 1 Descriptive statistics for demographic variables of Multiple Sclerosis patients for Study 1 and Study 2

Variable	Study 1 ($n = 2{,}952$)	Study 2 ($n = 1{,}767$)
Sex		
Male	423 (14.33%)	303 (17.15%)
Female	2,031 (68.80%)	1,464 (82.86%)
Age, mean (SD)	40.82 (9.35)	45.56 (9.31)
Time since diagnosis in years, mean (SD)	NA	3.69 (2.12)
Newly diagnosed		
Yes	1,054 (35.70%)	NA
No/unknown	1,898 (64.30%)	
Relapse		
Yes (T1)	NA	608 (34.41%)
Yes (T2)	NA	565 (31.98%)
Yes (T3)	NA	555 (31.41%)
Symptom change		
T1	NA	3.64 (1.15)
T2	NA	3.61 (1.08)
T3	NA	3.64 (1.04)
Group membership		
Stable	771 (26.12%)	NA
Actively relapsing	1,552 (52.57%)	NA
Progressing without relapsing	629 (21.31%)	NA

Table 2 Means and standard deviations for Performance Scale items (Study 1) and SF-12 scales (Study 2)

Measurement Occasions & Group Membership	Mobility	Hand Function	Fatigue	Cognitive	Bladder/Bowel	Sensory	Spasticity
Study 1—Performance Scales: higher scores indicate more disability							
Time 1							
Relapsing	1.52 (1.45)	1.22 (1.08)	2.58 (1.31)	1.62 (1.19)	1.18 (1.08)	1.89 (1.19)	1.51 (1.24)
Progressive	1.67 (1.47)	1.02 (0.98)	2.39 (1.26)	1.40 (1.17)	1.20 (1.04)	1.65 (1.17)	1.34 (1.22)
Stable	0.82 (1.17)	0.63 (0.86)	1.68 (1.22)	0.99 (1.01)	0.72 (0.90)	1.31 (1.03)	0.82 (1.05)
Time 2							
Relapsing	1.65 (1.50)	1.25 (1.09)	2.68 (1.31)	1.74 (1.23)	1.29 (1.15)	1.83 (1.22)	1.60 (1.29)
Progressive	1.79 (1.56)	1.10 (1.05)	2.45 (1.31)	1.45 (1.12)	1.24 (1.05)	1.59 (1.13)	1.48 (1.27)
Stable	0.78 (1.21)	0.66 (0.89)	1.70 (1.25)	1.02 (0.97)	0.75 (0.91)	1.13 (0.98)	0.77 (0.96)
Time 3							
Relapsing	1.75 (1.55)	1.31 (1.13)	2.70 (1.32)	1.77 (1.21)	1.37 (1.67)	1.84 (1.24)	1.65 (1.28)
Progressive	1.91 (1.67)	1.16 (1.07)	2.52 (1.28)	1.52 (1.16)	1.35 (1.10)	1.65 (1.16)	1.50 (1.29)
Stable	0.81 (1.24)	0.67 (0.88)	1.63 (1.25)	1.02 (1.25)	0.76 (0.91)	1.06 (0.96)	0.82 (0.99)

Measurement Occasions	PF	RF	BP	GH	VT	SF	RE	MH
Study 2—SF-12: higher scores indicate better health								
Time 1	6.80 (2.34)	6.26 (2.47)	5.40 (2.52)	4.30 (1.93)	3.13 (2.04)	7.17 (2.38)	7.42 (2.28)	5.87 (1.73)
Time 2	6.86 (2.17)	6.34 (2.32)	5.41 (2.54)	4.37 (1.95)	3.16 (1.96)	7.03 (2.43)	6.98 (2.38)	5.81 (1.65)
Time 3	6.70 (2.44)	6.23 (2.50)	5.52 (2.68)	4.32 (1.93)	3.05 (2.07)	7.13 (2.43)	7.50 (2.31)	5.92 (1.72)

Sample sizes in Study 1 are relapsing = 1,552, progressive = 629, and stable = 771. Total sample size = 2,952. Total sample size in Study 2 = 1,767

PF Physical Functioning, *RF* Role Functioning, *BP* Bodily Pain, *GH* General Health, *VT* Vitality, *SF* Social Functioning, *RE* Role Emotional, *MH* Mental Health

SF-12 means are sums of the items for each subscale and are not scaled to the standard 0–100 for computational convenience

Table 3 Overall goodness-of-fit and chi-square difference test results for Study 1 and Study 2

		CHISQ (*df*)	RMSEA (90% CI)	ECVI (90% CI)	Comparison models	CHISQ DIFF (*df*)	*P*	ECVI DIFF (90% CI)	CFI	TLI
Study 1—Performance Scales										
Models–Study 1										
1.1.1	Unidimensional measurement model	1,675.4 (495)	0.049 (0.047; 0.052)	0.75 (0.71; 0.79)					0.99	0.99
1.1.2	2 factor measurement model	899.2 (414)	0.035 (0.031; 0.028)	0.54 (0.51; 0.57)					1.00	0.99
1.2.1	Across occasion constraints on factor loadings and intercepts	1,281.6 (534)	0.038 (0.035; 0.040)	0.59 (0.55; 0.63)	1.2.1 vs. 1.1.2	382.4 (120)	<0.0001	0.04 (0.03; 0.07)	0.99	0.99
1.2.2	Sensory intercept, time 1 free, stable group	1,224.8 (533)	0.036 (0.034; 0.039)	0.57 (0.51; 0.58)	1.2.2 vs 1.2.1	56.8 (1)	<0.0001	0.02 (0.01; 0.03)	0.99	0.99
1.2.3	Sensory intercept, time 1 free, relapse group	1,150.6 (532)	0.034 (0.032; 0.037)	0.54 (0.49; 0.56)	2.2.3 vs 1.2.2	74.2 (1)	<0.0001	0.03 (0.02; 0.04)	1.00	0.99
1.3.1	As Model 1.2.2 but with addition of exogenous variables	1,461.8 (757)	0.031 (0.028; 0.033)	0.75 (0.69; 0.76)					0.99	0.99
1.3.2	Time lag restricted model	1,579.1 (808)	0.031 (0.029; 0.033)	0.76 (0.72; 0.80)	1.3.2 vs. 1.3.1	114.1 (50)	<0.0001	<0.01 (−0.01; 0.01)	0.99	0.99
Study 2—SF-12										
Models–study 2										
2.1.1	Measurement model from manual	2,305.3 (213)	0.076 (0.073; 0.079)	1.47 (1.37; 1.55)					0.94	0.93
2.1.2	Modified measurement model	1,147.1 (192)	0.053 (0.050; 0.059)	0.80 (0.73; 0.85)	2.1.2 vs. 2.1.1	1,158.2 (21)	<0.0001	0.63 (0.57; 0.70)	0.97	0.96
2.2.1	Across occasion constraints on factor loadings and intercepts	1,303.8 (218)	0.053 (0.050; 0.056)	0.86 (0.80; 0.93)	2.2.1 vs. 2.1.2	156.7 (26)	<0.0001	0.06 (0.04; 0.08)	0.97	0.96
2.2.2	Factor loadings and intercepts of RE free	1,224.8 (214)	0.052 (0.049; 0.055)	0.82 (0.76; 0.88)	2.2.2 vs. 2.2.1	79.0 (4)	<0.0001	−0.67 (−0.70; −0.62)	0.97	0.96
2.3.1	As Model 2.2.2 but with addition of exogenous variables	1,653.7 (376)	0.044 (0.042; 0.046)	1.188 (1.118; 1.262)					0.97	0.95
2.3.2	Free direct effects of age on MH	1,595.8 (373)	0.043 (0.041; 0.045)	1.159 (0.090; 1.231)	2.3.2 vs. 2.3.1	58.0 (3)	<0.0001	0.03 (0.02; 0.05)	0.97	0.96
2.3.3	Free direct effects of age on VT	1,555.0 (370)	0.043 (0.040; 0.045)	1.139 (1.072; 1.211)	2.3.3 vs. 3.3.2	40.8 (3)	<0.0001	0.02 (0.01; 0.03)	0.97	0.96

df degrees of freedom, *RMSEA* root mean square error of approximation, *ECVI* expected cross-validation index

Study 1, n = 2,952; Study 2, n = 1,767; Chi square refers to the normal theory weighted least squares chi-square

Models are numbered as follows: 1st number = study number, 2nd number = step in which model was computed, and 3rd number = possible modifications the relapse group at the first measurement occasion (χ^2diff(1) = 74.2, $P < 0.001$). No further model modifications were necessary. The ECVI difference tests were in agreement with these modifications (Table 3).

As can be seen in Fig. 1a, at the second and third measurement occasions, the intercepts of ‘sensory’ appear to decrease for the stable (0.37–0.18) and relapsing groups (0.37–0.18) relative to their Visible Disability (Fig. 1b). This response shift can be interpreted as recalibration and suggests that for these groups, when overall disability increases over time, specific sensory disability did not increase as much.

Step 3: Testing measurement invariance with respect to exogenous variables

In the final step, we used Model 1.2.2 and included additional exogenous variables (Table 3, Model 1.3.1). In all groups, at all occasions, we found positive correlations between age and both disability dimensions (see Table 4). The other exogenous variables, ‘sex’, ‘newly diagnosed’, ‘time between measurement occasions 1 and 2’, and ‘time between measurement occasions 2 and 3’ had correlations less than 0.1 with disability. When inspecting the SEPCs, we find that none were above our cut point of 0.10. Therefore, we concluded that there was no bias in the Performance Scales items with respect to these variables. See Fig. 1a and Table 4 for final model estimates.

For each group, the estimates of the common factor means of this final model (Model 1.3.2) are plotted against time in Fig. 1b and c. We see deterioration of progressive and relapsing patients (increasing visible disability scores) but no change in stable patients. There is little change within groups for internal disability, stable patients have the lowest internal disability means, and relapsing patients have the highest internal disability means.

Study 1 conclusion

Only “sensory disability” showed response shift. Considering the impact of response shift and true change on observed change in ‘sensory disability’, we see that an observed change of −0.19 for stable patients is almost fully attributable to response shift (−0.22), leaving only 0.03 of

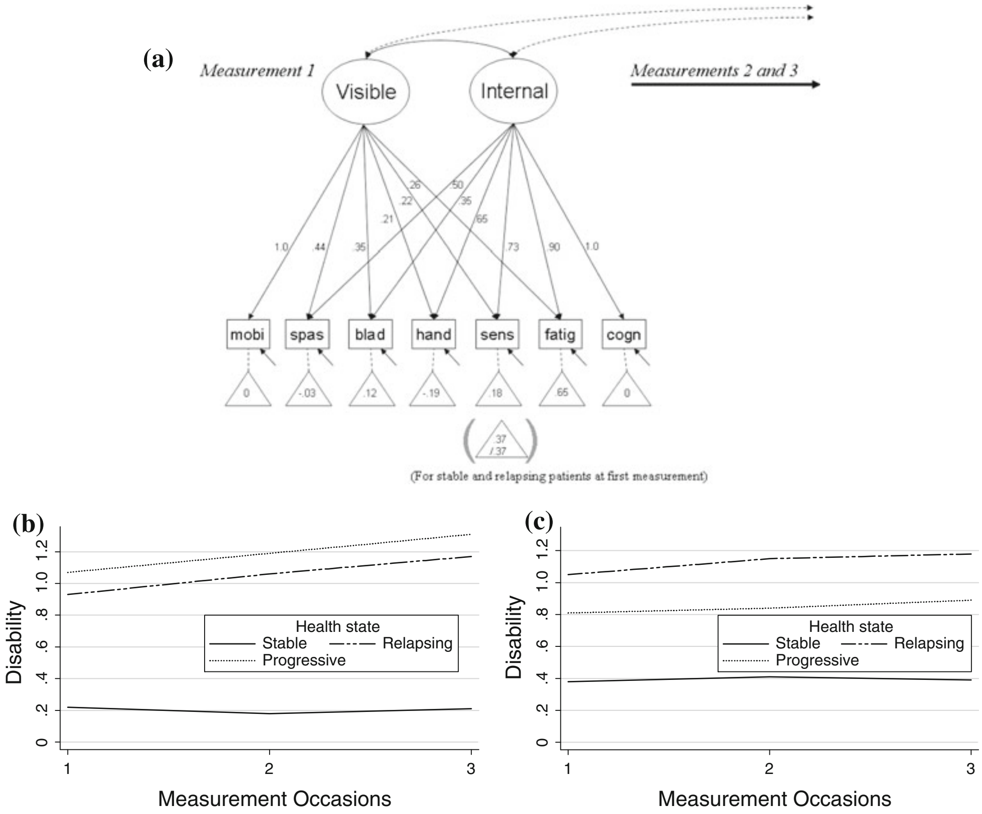

Fig. 1 **a** Performance Scale Measurement model at one measurement occasion. Factor loadings and intercepts of Model 1.2.2. *mobi* Mobility, *spas* Spasticity, *blad* Bladder/Bowel, *hand* Hand Function, *sens* Sensory, *fatig* Fatigue, *cogn* Cognitive. **b** Visible Disability Mean Change by Group and Between Models. **c** Internal Disability Mean Change by Group and Between Groups. *Note* Scaling of vertical axis is in standard deviations of the first measurement occasion common factor standard deviations

Table 4 Final model covariance and residual variance estimates

	Visible–T1	Internal–T1	Visible–T2	Internal–T2	Visible–T3	Internal–T3
Variance/covariances						
Visible–T1						
Stable	1.10					
Progressive	1.17					
Relapsing	1.52					
Internal–T1						
Stable	0.38	0.52				
Progressive	0.35	0.60				
Relapsing	0.56	0.73				
Visible–T2			1.26			
Stable	0.99	0.36	1.98			
Progressive	1.58	0.35	1.82			
Relapsing	1.43	0.56				
Internal–T2						
Stable	0.35	0.40	0.42	0.47		
Progressive	0.28	0.52	0.37	0.67		
Relapsing	0.50	0.63	0.65	0.77		
Visible–T3						
Stable	0.97	0.34	1.20	0.39	1.22	
Progressive	1.65	0.26	2.05	0.25	2.29	
Relapsing	1.41	0.52	1.77	0.58	1.93	
Internal–T3						
Stable	0.32	0.39	0.38	0.42	0.40	0.48
Progressive	0.25	0.52	0.30	0.61	0.35	0.69
Relapsing	0.51	0.61	0.64	0.69	0.65	0.79
Sex						
Stable	−0.01	0.01	−0.02	0.01	−0.03	0.01
Progressive	−0.09	<0.01	−0.08	−0.01	−0.09	−0.01
Relapsing	−0.08	−0.02	−0.07	−0.01	−0.08	−0.01
Age						
Stable	0.32	0.07	0.34	0.09	0.35	0.08
Progressive	0.43	0.04	0.50	0.02	0.51	0.01
Relapsing	0.37	0.12	0.43	0.14	0.43	0.13
Newly diagnosed	−0.26	−0.12	−0.26	−0.12	−0.26	−0.12
Time between T1 and T2			0.10	0.01	0.10	0.01
Time between T2 and T3					0.10	0.01

Residual variances	Mobility	Spasticity	Bladder/bowel	Hand function	Sensory	Fatigue	Cognitive
Stable							
T1	0.27	0.59	0.53	0.40	0.61	0.74	0.48
T2	0.23	0.45	0.51	0.42	0.55	0.80	0.48
T3	0.30	0.43	0.51	0.38	0.51	0.88	0.51
Progressive							
T1	0.41	0.82	0.77	0.60	0.82	0.79	0.75
T2	0.44	0.87	0.70	0.61	0.72	0.80	0.64
T3	0.40	0.81	0.80	0.60	0.76	0.80	0.67

Table 4 continued

Residual variances	Mobility	Spasticity	Bladder/bowel	Hand function	Sensory	Fatigue	Cognitive
Stable							
T1	0.57	0.83	0.76	0.58	0.76	0.77	0.72
T2	0.40	0.83	0.79	0.58	0.79	0.74	0.73
T3	0.43	0.76	0.86	0.61	0.78	0.73	0.67

true change. In the relapsing group, once we accounted for a response shift of −0.18, we see that an observed change of −0.06 underestimates a true change of 0.12. Still, we note that these effect sizes should be considered “small”.

Study 2 results

Participants were included if they had at least six of the 12 SF-12 items completed at three consecutive measurement occasions that were 6 (±3) months apart. This yielded a final sample of 1,767 patients. See Table 1 for sample characteristics and Table 2 for SF-12 observed scale means.

Step 1: Establishing a measurement model

The SF-12 PF, RP, BP, and GH scales are associated with PHYS, and VT, SF, RE, and MH are associated with MENT [8]. When replicating this structure, this model had only marginally satisfactory fit (Table 3, Model 2.1.1). Three sources of misfit were, at all measurement occasions, as follows: 1) covariances between residuals of PF and RP (χ^2diff(9) = 549.6, $P < 0.001$), 2) covariances between

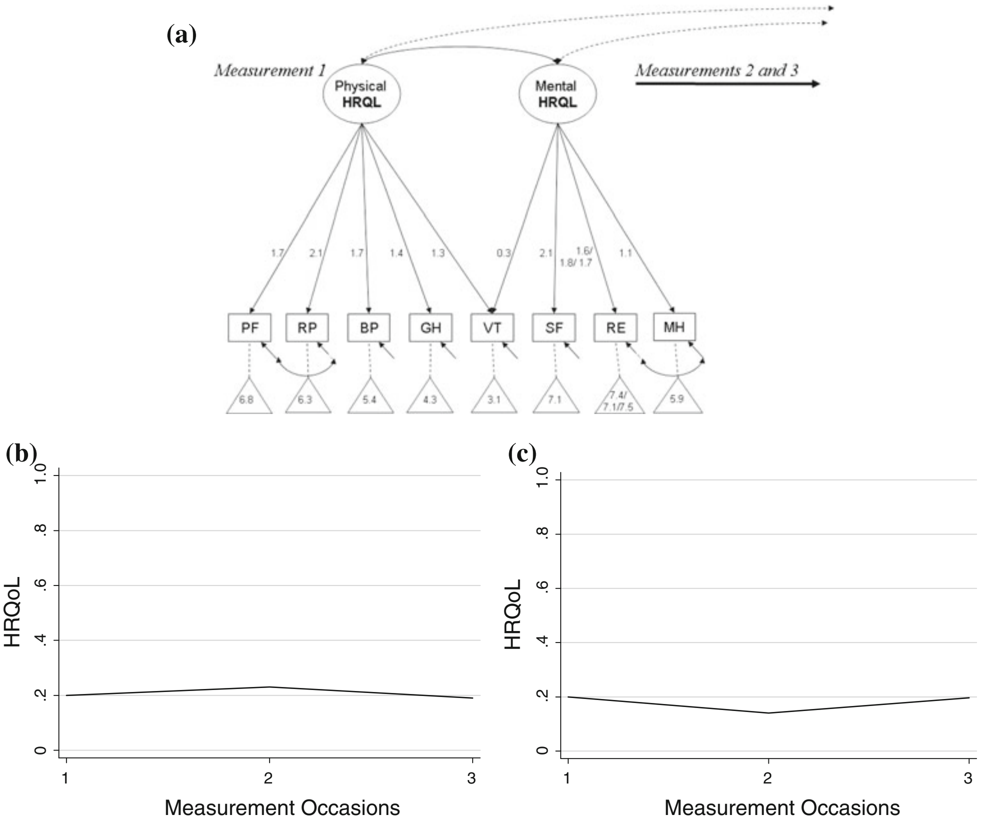

Fig. 2 **a** SF-12 Measurement model at one measurement occasion. Factor loadings and intercepts of Model 2.2.2. *PF* Physical Functioning, *RF* Role Functioning, *BP* Bodily Pain, *GH* General Health, *VT* Vitality, *SF* Social Functioning, *RE* Role Emotional, *MH* Mental Health. **b** PHYS HRQoL Mean Change Between Models. **c** MENT HRQoL Mean Change Between Models. *Note* Scaling of vertical axis is in standard deviations of the first measurement occasion common factor standard deviations

residuals of MH and RE (χ^2diff(9) = 452.5, $P < 0.001$), and 3) cross-loadings of VT on PHYS (χ^2diff(3) = 156.0.2, $P < 0.001$). These modifications produced a measurement model with satisfactory fit (Table 3, Model 2.1.2).

Step 2: Testing measurement invariance across measurement occasions

The equality constraints imposed in this step led to significantly deteriorated fit, suggesting the presence of response shift (Table 3, Model 2.2.1). The global test associated with the scale RE resulted in the largest OPCs and a significant chi-square difference test. Therefore, the parameters associated with RE were free to be estimated (Fig. 2a). As can be seen in Fig. 2a, the intercept of RE at the second measurement occasion was lower (7.1) than at the first and third measurement occasions (7.4 and 7.5). This suggests that there is uniform recalibration for the RE scale: Patients seemed less inclined to report high RE at the second measurement occasion as compared to the first and third measurement occasions, given similar MENT HRQoL. No further response shifts were found.

Step 3: Testing measurement invariance with respect to exogenous variables

Adding additional exogenous variables to Model 2.2.2 resulted in a satisfactorily fitting model (Table 3, Model 2.3.1). We found large negative correlations of age and relapse with PHYS and MENT and positive correlations of symptom change with PHYS and MENT. The correlations between sex and time since diagnosis and PHYS and MENT are considered very small (<0.01) (Table 5). With Model 2.3.1 as the comparison model, we proceeded to test for bias with respect to the exogenous variables using the global tests and OPCs. Significant direct effects of age on MH (Model 2.3.2) were found. We also found significant direct effects of age on VT (Model 2.3.3). In the next iteration, no further significant effects were found.

We tested whether the measurement bias in MH and VT with respect to age was consistent across measurement occasions. As the inclusion of equality constraints across measurement occasions did not worsen model fit (χ^2diff (4) = 7.80, $P = 0.099$), we concluded that the bias is consistent and did not indicate response shift. Given the negative correlations between age and PHYS and MENT (Table 5),

Table 5 Final model variance/covariances and residual variance estimates

	PHYS HRQoL –T1	MENT HRQoL –T1	PHYS HRQoL –T2	MENT HRQoL –T2	PHYS HRQoL –T3	MENT HRQoL –T3
Variance/Covariances						
PHYS HRQoL –T1	1					
MENT HRQoL –T1	0.87	1				
PHYS HRQoL –T2	0.87	0.78	0.94			
MENT HRQoL –T2	0.81	0.81	0.84	0.95		
PHYS HRQoL –T3	0.91	0.80	0.90	0.83	1.05	
MENT HRQoL –T3	0.77	0.78	0.77	0.81	0.91	0.98
Sex	−0.03	0.002	−0.01	−0.01	−0.02	−0.005
Age	−0.25	−0.12	−0.21	−0.13	−0.26	−0.13
Time since diagnosis	−0.04	0.02	0.01	−0.05	−0.02	−0.03
Symptom change–T1	0.63	0.52	0.52	0.48	0.52	0.39
Symptom change–T2	0.52	0.42	0.61	0.53	0.57	0.41
Symptom change–T3	0.41	0.36	0.42	0.39	0.60	0.49
Relapse–T1	−0.15	−0.15	−0.13	−0.12	−0.13	−0.12
Relapse–T2	−0.13	−0.13	−0.15	−0.14	−0.14	−0.12
Relapse–T3	−0.12	−0.11	−0.13	−0.11	−0.17	−0.14

	PF	RP	BP	GH	VT	SF	RE	MH
Residual variances								
T1	2.27	1.66	3.24	1.68	1.97	1.31	2.65	1.76
T2	1.99	1.62	3.28	1.65	1.82	1.56	2.61	1.67
T3	2.51	1.60	3.91	1.64	2.02	1.62	2.59	1.79

the bias on MH (0.20) and VT (0.19) with respect to age suggests that older patients reported better MH and VT than was expected. The estimates of the common factor means of this model (Model 2.3.3) did not show any change (Fig. 2b and c).

Study 2 conclusion

When we consider the impact of response shift and true change on observed change in EF, we see that the observed change of 0.23 is almost fully attributable to response shift (0.17), leaving only 0.06 of true change. However, only on the second measurement occasion, we found an indication of response shift in EF, which hinders substantive interpretation. So we concluded that this may be a chance finding.

Discussion

We have illustrated two different ways in which SEM can be used to investigate response shift. With the present data, we found uniform recalibration response shift in the sensory disability item of the Performance Scales (Study 1), indicating that stable and relapsing patients initially overestimate their sensory disability. Apparently, in comparison with their general performance disability, they initially worry more about their sensory disability but then become accustomed to their situation, whereas progressive patients continue to deteriorate in their performance disability and, as a result, do not become accustomed to sensory disability. In a study investigating progressive MS patients, it was shown that the presence of sensory disability led to an increased length of time to reach a severe level of disability [20]. In another study comparing progressive and relapsing patients, it was found that relapsing patients had higher initial sensory disability than progressive patients [21]. It may be possible that the gradual progression of disease seen in progressive patients leads to a slight worsening of sensory disability over time, which is difficult to become accustomed to.

We did not find clearly interpretable response shift in the SF-12 (Study 2), nor did we find any change in HRQoL. However, two measurement biases that are not response shift, as they are constant across measurement occasions, were found: age on MH and age on VT. The correlations between age and PHYS and MENT HRQoL are negative; however, the direct effects of age on MH and VT are positive. This suggests that increased age affects MH and VT in a different way than would be expected due to the correlations between the age and the common factors.

A possible explanation for the limited response shift findings is that the NARCOMS registry patients are not subjected to a planned intervention, so there is no clear catalyst of health state changes, other than self-reported relapse and symptoms. Therefore, the sizes of the response shifts found are small, and accounting for these response shifts does not cause large effects on the mean change in performance disability or HRQoL. Though importantly, the response shifts do change the interpretation of the observed changes. In Study 1 for the stable patients and in Study 2 for all patients, the response shift accounts for essentially all observed change, leaving essentially no true change. With the relapsing patients in Study 1, the observed change is underestimated, and after taking response shift into consideration, true change appears small and negative.

These two studies highlight how SEM can be used to detect measurement bias under different circumstances. The steps used are hierarchical; however, there is the flexibility (1) to account for health state by splitting the sample into subsamples or by including exogenous variables, (2) to use time-varying or time-constant exogenous variables, (3) to use different search strategies for detecting response shift, (4) inclusion of fixed or random measurement occasions and, not discussed in this paper, (5) to investigate different longitudinal structures like autoregressive and latent growth curve structures [22]. Some of these decisions are made based on the design of the study or sample size available, and for each decision made there are trade-offs. A persistent problem is that the decision of when to stop investigating measurement bias is relatively subjective. Because of the large number of tests to consider, despite our strict criteria for what we consider response shift, we still need to guard against chance findings [18]. Still, SEM offers a useful statistical approach for response shift detection as it can be tailored to the specifics of the study design.

Acknowledgments This work was funded in part by a Visiting Scientist Fellowship from the Consortium of Multiple Sclerosis Centers Foundation to Dr. Schwartz.

Appendix 1

Study 1

Step 1: Establishing a measurement model

The Performance Scales were originally reported to be a unidimensional construct [7]. If the confirmatory factor analysis (CFA) used to assess the longitudinal multigroup uni-dimensional model does not fit, then exploratory factor analysis is used to determine an alternative model, and the fit of this model is assessed using CFA.

Maximum likelihood estimation is used for parameter estimation. We assess the overall fit of our models with the chi-square test of exact fit, the root mean square error of approximation (RMSEA) [12], the expected cross-validation index (ECVI) [12], the comparative fit index (CFI) [10], and the Tucker Lewis index (TLI) [11]. Emphasis, however, is placed on the RMSEA and ECVI fit statistics as confidence intervals can be calculated for these fit statistics. A non-significant chi-square value indicates good fit. As it is sensitive to small deviations between model and data, especially when the sample size is large, we also consider approximate fit indices that relax the stringent requirement on chi-square that the model has exact fit to the population. RMSEA $< .08$ indicates satisfactory fit; RMSEA $< .05$ indicates close fit. ECVI cannot be used as a stand-alone index but can be used to compare alternative models; a smaller ECVI value suggests better model fit [12]. Finally, both the CFI and the TLI assess the improvement in fit from a null model with no relationships assumed between variables. Values >.95 for the CFI and >.90 for the TLI indicate reasonable fit of the model to data.

If a new model is specified due to misfit as indicated by modification indices, standardized residuals or standardized expected parameter changes (SEPC), then the change in overall model fit is assessed with the chi-square difference test and the ECVI difference test [13]. The chi-square difference test is used to assess whether the alternative model fit is significantly better than the fit of the null model. A significant result indicates that the alternative model has better fit than the null model. The ECVI difference is the difference between the ECVI values of the null and alternative models. If the confidence interval of this test does not include zero, then the change is significant. The ECVI test is based on the chi-square test but it penalizes models containing more free parameters [13].

Step 2: Testing invariance across measurement occasions

In this step, we take the final model of Step 1 and simultaneously constrain all factor loadings and intercepts to be equal across measurement occasions and groups. Across-occasion invariance (lack of measurement bias) of the factor loadings and intercepts is assessed by comparing this model with the final model of Step 1 using the chi-square difference test. A significant test provides strong evidence that response shift is present as it is possible that the equality constraints imposed are not tenable. However, if the test is not significant, it may still be possible that one of the equality constraints is not tenable. Therefore, we still search for bias, as a single, yet substantially important response shift may not cause significant deterioration in the overall model fit.

To detect measurement bias, we examine modification indices and SEPCs [13]. One prerequisite for meaningful model respecification is that large modification indices are substantively interpretable; however, this alone does not ensure a substantial change in the parameter estimate, especially in large samples. Therefore, we also consider the associated SEPC. If both are large, then there is a significant improvement in the overall model and substantial change in the parameter estimate. As there are a large number of modification indices to consider, we stop investigating modification indices when none are greater than 12.83 [14]. This critical value has been adjusted for the number of tests in consideration so as to maintain a family-wise type 1 error rate of 5%. For the SEPCs, we consider >0.10 significant [15]. The impact of any response shift found is assessed by using Oort's [2] partitioning formula to evaluate the contribution of response shift and true change in terms of Cohen's effect size d [16].

Step 3: Testing invariance with respect to exogenous variables

Using the final model in Step 2, we now include 'age', 'sex', 'newly diagnosed', 'time between measurement occasions 1 and 2', and 'time between measurement occasions 2 and 3' as additional exogenous variables. These five exogenous variables correlate with each other and with the common factors, but their relationship with the observed items should be fully explained by these correlations. If large modification indices and SEPCs are present between the exogenous variables and the observed items, this is an indication of bias and requires the estimation of direct effects. If the estimates of the direct effects change over time (i.e., cannot be constrained to be equal across measurement occasions), we interpret this as response shift. However, if the direct effects do not change over time, we consider this measurement bias. Large modification indices and SEPCs will be evaluated using the same criteria as outlined in Step 2.

Appendix 2

Study 2

Step 1: Establishing a measurement model

The first goal is to find a satisfactory measurement model for the SF-12. We begin with the measurement model comprising two common factors: PHYS and MENT HRQoL. If this model does not fit, we use modification indices and standardized residuals [13, 18] to identify misspecification and develop an alternative model. As we are evaluating the measurement model longitudinally, we require the same observed variables to be associated with the same common factors across measurement occasions. If the model is modified, the changes are assessed using the chi-square difference test and ECVI difference test as explained in Study 1. Overall model fit is assessed using the same statistics as used in Study 1.

Step 2: Testing invariance across measurement occasions

Using the final measurement model from Step 1, all factor loadings and intercepts of the final model of Step 1 are simultaneously constrained to be equal across measurement occasions, like in Study 1. The assessment of overall model fit and change in model fit compared to the final model of Step 1 are again done in the same way as in Study 1.

To detect response shift, we use a different search strategy from Study 1 where we test individual constraints. Here, we follow the procedure outlined in King-Kallimanis et al. [19], where all observed scales are tested with a smaller number of global tests that free multiple constraints simultaneously. In this study, we use eight global tests, one for each observed scale. That is, for each of the eight scales, the equality constraints on both the factor loadings and the intercepts at all three measurement occasions are removed. The fit of each of these eight new models is compared to the fully constrained model using the chi-square difference test. The impact of freeing the parameters on the estimated parameter values is assessed by calculating the observed parameter changes (OPC). The OPCs are scaled for ease of comparison, and they are the actual difference between the standardized factor loadings and intercepts of the null model, and the standardized factor loadings and intercepts of the altered model. Corresponding to Cohen's small effect sizes, we consider an OPC indicating a difference of 0.1 between factor loadings or 0.2 between intercepts to be large [16]. We consider both values because small deviations in the observed and expected covariance matrix may lead to significant model improvement, but not substantial parameter change.

After running the eight tests, the model specifics are checked. In particular, scales with OPCs that meet our criteria that are in conjunction with a significant chi-square difference test are considered as exhibiting response shift. The factor loadings and intercepts of this scale remain unconstrained, and the remaining scales are retested with an adjusted significance level. We continue iteratively retesting the remaining scales, until no large OPC with significant chi-square difference test is found.

Although there are fewer tests to consider when using the global tests, when the number of iterations increases, so does the number of tests. Therefore, when considering the significance of the chi-square difference test, we use a Bonferroni-adjusted level of significance, with the family-wise level of significance at 5% divided by the number of tests under consideration for this particular step of the analysis [14].

Step 3: Testing invariance with respect to exogenous variables

We extend the final model of Step 2 to include 'age', 'sex', 'time since diagnosis', 'relapse in the past 6 months', and 'symptom change in the last 6 months' as exogenous variables. We hypothesize that these variables have the potential to induce bias on the observed scales. These additional variables are free to correlate with each other and with the common factors; however, all direct effects between the observed scales are fixed to zero. To test for response shift, we fit new models where the direct effects of the exogenous variables are free to be estimated. For example, for 'sex' we fit eight new models, with the effect of sex on an observed scale for three measurement occasions. This results in three additional parameters to be estimated. The impact of these direct effects is assessed with OPCs and the chi-square difference test. If the largest effects meet our criteria like in Step 2, we leave these parameters free to be estimated and start the process over again and stop when no freed direct effects meet our criteria. Once any biases have been accounted for, this final model can be used to assess true change in the attribute of interest using the same partitioning formula we use in Study 1.

References

1. Sprangers, M. A. G., & Schwartz, C. E. (1999). Integrating response shift into health-related quality of life research: a theoretical model. *Social Science and Medicine, 48*, 1507–1515.
2. Oort, F. J. (2005). Using structural equation modeling to detect response shifts and true change. *Quality of Life Research, 14*, 587–598.
3. Visser, M. R. M., Oort, F. J., & Sprangers, M. A. G. (2005). Methods to detect response shift in quality of life data: A convergent validity study. *Quality of Life Research, 14*, 629–639.
4. Oort, F. J., Visser, M. R., & Sprangers, M. A. (2009). Formal definitions of measurement bias and explanation bias clarify measurement and conceptual perspectives on response shift. *Journal of Clinical Epidemiolgy, 62*, 1126–1137.
5. Oort, F. J. (2005). Towards a formal definition of response shift (In reply to G.W. Donaldson). *Quality of Life Research, 14*, 2353–2355.
6. Schwartz, C.E., Sprangers, M.A.G., & Vollmer, T. (2010). A rashomon approach to detecting response shift in patients with multiple sclerosis: a head-to-head comparison of four statistical techniques. *Quality of Life Research*, Current Issue.

7. Schwartz, C. E., Vollmer, T., & Lee, H. (1999). Reliability and validity of two self-report measures of impairment and disability for MS. *Neurology, 52*, 63–70.
8. Ware, J. E., Kosinski, M., & Keller, S. D. (1996). A 12-item short-form health survey—Construction of scales and preliminary tests of reliability and validity. *Medical Care, 34*, 220–233.
9. Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via Em Algorithm. *Journal of the Royal Statistical Society Series B-Methodological, 39*, 1–38.
10. Bentler, P. M. (1990). Comparative Fit Indexes in Structural Models. *Psychological Bulletin, 107*, 238–246.
11. Tucker, L. R., & Lewis, C. (1973). The reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 37*, 1–10.
12. Browne, M. W., & Cudeck, R. (1992). Alternative Ways of Assessing Model Fit. *Sociological Methods & Research, 21*, 230–258.
13. Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York: Wiley.
14. Tabachnick, B. G., & Fidel, L. S. (2006). *Using Multivariate Statistics. Allyn & Bacon, Inc, Needham Heights*. USA: MA.
15. Saris, W. E., Satorra, A., & van der Veld, W. M. (2009). Testing Structural Equation Models or Detection of Misspecifications? *Structural Equation Modeling-A Multidisciplinary Journal, 16*, 561–582.
16. Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawerence Erlbaum associates.
17. Neale, M. C. (2010). *MxGui*. Richmond, VA: In. VCU.
18. Jöreskog, K. G., & Sörbom, D. L. I. S. R. E. L. (1996). *6 user's guide* (Vol. 2). Chicago, IL: Scientific Software International, Inc.
19. King-Kallimanis, B. L., Oort, F. J., & Garst, G. J. A. (2010). Using structural equation modelling to detect measurement bias and response shift in longitudinal data. *Asta-Advances in Statistical Analysis, 94*, 139–156.
20. Koch, M., Kingwell, E., Rieckmann, P., & Tremlett, H. (2009). The natural history of primary progressive multiple sclerosis. *Neurology, 73*, 1996–2002.
21. Sola, P., Mandrioli, J., Simone, A.M., Ferraro, D., Bedin, R., Annecca, R., et al. (2010). Primary progressive versus relapsing-onset multiple sclerosis: presence and prognostic value of cerebrospinal fluid oligoclonal IgM. *Multiple Sclerosis,* doi: 10.1177/1352458510386996.
22. Oort, F. J. (2001). Three-mode models for multivariate longitudinal data. *British Journal of Mathematical and Statistical Psychology, 54*, 49–78.