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

Verdam, M.G.E.

Publication date
2017

Document Version
Other version

License
Other

CHAPTER 7

Item Bias Detection in the Hospital Anxiety and Depression Scale Using Structural Equation Modeling: Comparison with Other Item Bias Detection Methods

Comparison of patient reported outcomes may be invalidated by the occurrence of item bias, also known as differential item functioning. We show two ways of using structural equation modeling (SEM) to detect item bias: (1) multigroup SEM, which enables the detection of both uniform and nonuniform bias, and (2) multidimensional SEM, which enables the investigation of item bias with respect to several variables simultaneously. We apply both SEM approaches for the detection of gender- and age-related bias in the items of the Hospital Anxiety and Depression Scale (HADS; Zigmond & Snaith, 1983) from a sample of 1,068 patients. Results were compared to the results of the ordinal logistic regression, item response theory, and contingency tables methods reported by Cameron, Scott, Adler and Reid (2014). Both SEM approaches identified two items with gender-related bias and two items with age-related bias in the Anxiety subscale, and four items with age-related bias in the Depression subscale. Results from the SEM approaches generally agreed with the results of Cameron et al., although the SEM approaches identified more items as biased. We conclude that SEM provides a flexible tool for the investigation of item bias in health-related questionnaires. Multidimensional SEM has practical and statistical advantages over Multigroup SEM, and over other item bias detection methods, as it enables item bias detection with respect to multiple variables, of various measurement levels, and with more statistical power, ultimately providing more valid comparisons of patients' well-being in both research and clinical practice.

This chapter is based on: Verdam, M. G. E., Oort, F. J., & Sprangers, M. A. G. (in press). Using structural equation modeling to detect item bias: Comparison with other item bias detection methods. *Quality of Life Research*.

Introduction

Comparing assessments of patient reported outcomes is important for clinical practice and research. However, such comparisons may be invalidated by the occurrence of differential item functioning (DIF). DIF, also referred to as item bias, occurs when two people with the same value on the trait of interest (e.g., well-being), have a different probability of giving a certain response on an item from a questionnaire or test that measures the trait of interest, due to differences on other variables (e.g., age, gender, attitudes, mood, treatment condition, etc.). Mellenbergh (1989) gave a formal definition of item bias: An item X measuring trait T is unbiased with respect to another variable V, if and only if:

$$f_1(X \mid V = v, T = t) = f_2(X \mid T = t), \quad (1)$$

where f_1 is the distribution of the item responses given the values v and t of variables V and T, and f_2 is the distribution of item responses given only the values t of variable T. Mellenbergh emphasized the generality of the definition, where the variables X, V and T may have nominal, ordinal or interval measurement scales. In the presence of item bias, differences between two people on observed item scores may not reflect 'true' differences on the trait variable (e.g., men and women may score differently on an item that measures well-being, even though their well-being does not differ). If the bias is uniform, it is consistent for all levels of the latent trait (e.g., the size of the bias is independent of the level of well-being). When the bias is nonuniform, it differs for different levels of the latent trait (e.g., the difference may be larger for higher levels of well-being).

Statistical methods for the detection of item bias can be distinguished based on their operationalization of the trait variable T. One group of methods use the summary of the observed item scores (i.e., the scale score) to operationalize the trait variable (e.g., loglinear models, contingency tables methods, logistic regression models, standardization methods), and another group of methods operationalize an unobserved latent trait variable (e.g., item response theory analysis and structural equation modelling methods) (Millsap & Everson, 1993). We further distinguish between methods that can detect uniform item bias, and methods that can also detect nonuniform item bias. Although advantages have been made to enable the investigation of nonuniform item bias, it is not always easily implemented and therefore not often applied.

Cameron, Scott, Adler and Reid (2014) recently investigated the equivalence of three different bias detection methods for the detection of gender- and age-related bias in the items of the Hospital Anxiety and Depression Scale (HADS; Zigmond & Snaith, 1983). They applied ordinal logistic regression, item response theory (IRT), and contingency tables methods to investigate item bias in the anxiety and depression subscales of the HADS separately. All three methods were used to detect uniform item bias only. Although Cameron et al. mention structural equation modeling (SEM) methods as a fourth option that can be applied to investigate item bias, they did not incorporate SEM methods in their comparison.

Compared to the three item bias detection methods applied by Cameron et al., SEM methods may have several important advantages. The multigroup SEM approach can be applied to detect bias in observed item scores with respect to group-membership (e.g., gender or age-category) and a continuous latent trait variable (e.g., depression or anxiety). Advantages of the SEM approach are that it uses a latent trait operationalization, it enables the detection of both uniform and nonuniform bias, and that possible item bias can be taken into account to assess true differences between groups. In addition, the flexibility of the SEM framework allows for extensions to multidimensional models. This enables the investigation of item bias with respect to any factor or variable (e.g., continuous or categorical, latent or manifest). Uniform bias can then be investigated by testing the significance of direct effects of these additional factors on the observed items. Advantages of the multidimensional SEM approach over the multigroup SEM approach are that continuous variables can be included in the model without categorizing them, and that item bias can be investigated with respect to several variables simultaneously.

The objective of the present paper is threefold. First, we apply the multigroup SEM approach to investigate gender- and age-related item bias in each subscale of the HADS. Second, we apply the multidimensional SEM approach to both subscales of the HADS, and investigate gender- and age-related item bias simultaneously. Third, we compare the results of both SEM approaches to the results of the three item bias detection methods that were investigated by Cameron et al (2014).

Method

A total of 1068 adults who consulted a primary care professional in North East Scotland completed the HADS (for more details on data collection see Cameron, Lawton, & Reid, 2009). The HADS is a 14-item self-report instrument that consists of an anxiety (HADS-A; 7 items) and depression (HADS-D; 7 items) subscale where higher scores represent greater symptom severity. All items are answered on an ordinal response scale with four response categories.

Statistical analyses

Structural equation modeling (SEM) was used to investigate gender- and age-related item bias in the anxiety and depression subscales of the HADS. To accommodate discrete ordinal responses, we applied the structural equation modeling approach proposed by Verdam, Oort, & Sprangers (2016). This approach includes two stages: (1) establishing a model of underlying continuous variables that represent the observed discrete variables, (2) using these underlying continuous variables to establish a common factor model for the detection of item bias, and to assess true change. The SEM approach with discrete data was originally illustrated with longitudinal data, but can also be applied to the multigroup situation. Statistical analyses were performed using the PRELIS (Stage 1) and LISREL (Stage 2) programs (Jöreskog & Sörbom, 1996).

Multigroup SEM procedure

Gender- and age-related item bias was investigated for the anxiety and depression subscales of the HADS separately, by comparing a 'reference' and 'focal' group. For age, there were 814 participants in the reference group (< 65 years) and 254 participants in the focal group (≥ 65 years). For gender, there were 633 participants in the reference group (women) and 435 in the focal group (men). Figure 1 gives a graphical representation of the multigroup model for item bias detection.

In Stage 1, the model of underlying continuous variables that represent the observed discrete variables was used to estimate thresholds and polychoric correlations under the assumption of bivariate normality in both groups. Thresholds of the same items were constrained to be equal across groups. The tenability of the assumption of underlying bivariate normality in each group was evaluated using the root mean square error of approximation (RMSEA; Steiger & Lind, 1980; Steiger, 1990), with the criterion that RMSEA values should not be larger than 0.1 (Jöreskog, 2002). When the hypothesis of bivariate normality under equal thresholds holds for all pairs of variables, the estimated polychoric correlations, variances and means of the underlying continuous variables can be used in subsequent analyses of Stage 2. When the hypothesis of bivariate normality does not hold, then this indicates that the assumption of multivariate normality (under equal thresholds) is not tenable. A possible solution for this problem is to eliminate the offending variable(s).

In Stage 2, *Step 1*, the estimates from the underlying variables from Stage 1 were used to establish a multigroup common factor model (e.g., a one-factor model for "Anxiety", with seven indicator items, for both men and women; see Figure 1). The Measurement Model has no across groups constraints. To test whether the Measurement Model holds, goodness-of-fit can be assessed using the chi-square test statistic. The chi-square test was used to evaluate exact goodness-of-fit, where a significant chi-square indicates a significant difference between data and model. In addition, the RMSEA value was used as a measure of approximate goodness-of-fit, where values below .08 indicate 'reasonable' approximate fit and below .05 'close' approximate fit (Browne & Cudeck, 1992).

In *Step 2*, the No Item Bias Model was fitted to the data, where all measurement parameters were constrained to be equal across groups. Item bias was operationalized as across-group differences between values of intercepts (i.e., uniform item bias; across-group differences in the endorsement of an item, independent from the latent trait variable) and differences between common factor loadings (i.e., nonuniform item bias; across-group differences in the extent to which an item measures the latent trait variable). To test for the presence of item bias, the No Item Bias Model can be compared to the Measurement Model. The chi-square difference test was used to test the difference in exact fit, where a significant chi-square difference indicates that the No Item Bias Model has significantly worse fit as compared to the Measurement Model. If the invariance restrictions of the No Item Bias Model led to a significant deterioration in model fit, this indicated the presence of item bias.

In case of item bias, in *Step 3*, a step-by-step modification of the No Item Bias Model was used to arrive at the Final Model in which all items that showed item bias were taken into account.

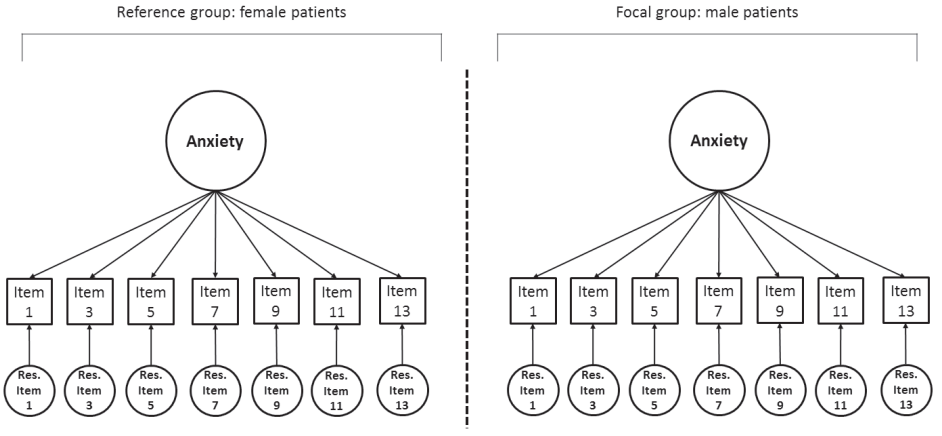

Figure 1 | The two-group model for gender-related item bias detection in the anxiety subscale of the HADS

Notes: Similar models have been used for the detection of age-related item bias in the anxiety subscale of the HADS, and for the detection of gender- and age-related item bias in the depression subscale of the HADS. The squares represent the underlying continuous variables associated with the observed item responses of Item 1 to Item 13. The circle at the top is the underlying common factor Anxiety, which represents everything that Item 1 to Item 13 have in common. Each item is associated with a residual factor, which represents everything that is specific to the corresponding item Item bias is operationalized as across group differences in intercepts (uniform) and factor loadings (nonuniform).

The identification of item bias was guided by an iterative procedure, where each across-group constraint was set free one at a time, and the freely estimated parameter that led to the largest improvement in fit was included in the model. Each indication of bias was tested by evaluating the improvement in model fit using the chi-square difference test to evaluate differences in exact fit. In addition, the Final Model was compared to the Measurement Model to test equivalence of exact fit as an indication that all apparent item bias was taken into account. To give an indication of the size of the detected item bias, we calculated Cohen's *d* effect size indices for the impact of both uniform and nonuniform item bias on the differences between the item means across groups (see Oort, 2005 for more details), where values of 0.2, 0.5, and 0.8 indicate small, medium, and large effects (Cohen, 1988). Following the example of Cameron et al. (2014), we used importance criteria in addition to significance criteria for the detected item bias (see also Table 2), where item bias was considered 'important' when the size of the item bias was larger than 0.2.

In *Step 4*, the estimates of common factor means of the Final Model, in which all apparent item bias was taken into account, was used to assess 'true' differences between the groups. Cohen's *d* effect size was calculated to give an indication of the size of the difference.

Multidimensional SEM procedure

Gender- and age-related item bias was investigated for the anxiety and depression subscales of the HADS simultaneously, by including both age and gender as exogenous variables in the multidimensional model. Figure 2 gives a graphical representation of the multidimensional model for item bias detection. The procedure for item bias detection using the multidimensional

approach was largely similar to the procedure for item bias detection using the multigroup approach. Here, we describe only the differences in the procedures. In Stage 1, correlations between all variables in the model (i.e., the underlying variables that correspond to the observed items and the exogenous variables) were estimated. In Stage 2, *Step 1*, the estimates from the underlying variables from Stage 1 were used to establish a multidimensional common factor model that included the common factors "Anxiety" and "Depression", each with seven indicator variables. In *Step 2*, the multidimensional model was extended to include the variables "Age" and "Gender". These variables were allowed to correlate with the common factors, but all direct effects of Age and Gender on the items were constrained to zero. This model is referred to as the No Item Bias Model. The overall model fit of this model was used to give an indication of the presence of item bias. In *Step 3*, an iterative procedure was used, where each constrained direct effect of the exogenous variables age and gender was set free to be estimated one at a time, and the freely estimated parameter that led to the largest improvement in fit was included in the model. When freeing additional parameters did not lead to an improvement in model fit, this was taken as an indication that all apparent bias was taken into account. The importance criterion for item bias was evaluated using the standardized direct effects, which can be interpreted as effect size *r*, with values of 0.1, 0.3, and 0.5 indicating small, medium, and large effect sizes (Cohen, 1988). In *Step 4*, the correlations between the exogenous variables age and gender and the common factors of the Final Model, in which all apparent bias has been taken into account, were used to assess 'true' differences between the genders, and 'true' associations with age.

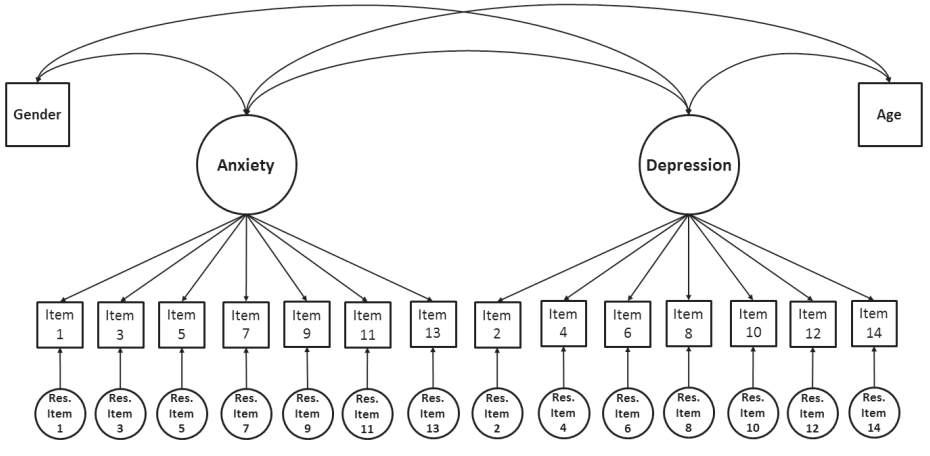

Figure 2 | The multidimensional model for gender- and age-related item bias detection in the anxiety and depression subscales of the HADS

Notes: The squares represent the underlying continuous variables associated with the observed item responses of Item 1 to Item 14. The circles at the top are the underlying common factors Anxiety and Depression. Anxiety represents everything that Item 1 to Item 13 have in common, whereas Depression represents everything that Item 2 to Item 14 have in common. Each item is associated with a residual factor, which represents everything that is specific to the corresponding item The multidimensional model includes two exogenous variables: Gender and Age. Uniform item bias is operationalized as significant direct effects of the exogenous variables on the indicator variables (i.e., Item 1 to Item 14).

Results

Results of item bias detection are presented in Table 1, where the items that were identified as having bias by one of the SEM approaches are compared to the results from Cameron et al. (2014) in Table 2. We first report results of the multigroup SEM approach, and then of the multidimensional SEM approach.

Multigroup SEM approach

Results of Stage 1 indicated that the hypothesis of bivariate normality under equal thresholds was tenable for all item pairs, for both subscales and both gender- and age-groups. Estimated polychoric correlations, variances and means were used in subsequent analyses of Stage 2. We report results of gender- and age-related item bias for each subscale of the HADS separately.

Anxiety subscale. *Gender related item bias.* Results of Stage 2 indicated that the Measurement Model showed close approximate fit (Model 1a, Table 1). Imposition of equality constraints on measurement parameters across groups yielded the No Item Bias Model (Model 1b). The No Item Bias Model showed a significant deterioration in model fit as compared to the Measurement Model, indicating the presence of gender-related item bias of the HADS-A (see Table 1). Indications of uniform bias were detected for Item 9 ($CHISQ_{diff}(1) = 14.54$, $p < .001$), and for Item 11 ($CHISQ_{diff}(1) = 30.57$, $p < .001$). The Final Model, in which both biases were incorporated in the model, showed close approximate fit (Model 1c, Table 1). Although the Final Model did not yield equivalent fit as compared to the Measurement Model, freeing additional parameters did not significantly improve model fit.

Age-related item bias. The Measurement Model showed close approximate fit (Model 3a, Table 1). The No Item Bias Model yielded a significant deterioration in model fit, indicating the presence of age-related item bias of the HADS-A. Two items with uniform bias, and one item with nonuniform bias were identified. The Final Model that incorporated these three biases (Model 3c) showed equivalent fit compared to the Measurement Model (see Table 1). Uniform bias was detected for Item 1 ($CHISQ_{diff}(1) = 18.36$, $p < .001$), and for Item 13 ($CHISQ_{diff}(1) = 50.78$, $p < .001$), whereas nonuniform bias was detected for Item 3 ($CHISQ_{diff}(1) = 12.31$, $p < .001$).

True differences between the groups. Inspection of common factor means showed that men score significantly lower on the Anxiety factor as compared to women ($d = -0.30$, $p < .001$), and that patients aged above 65 scored significantly lower on the Anxiety factor compared to patients aged below 65 ($d = -0.76$, $p < .001$).

The depression subscale. *Gender-related item bias.* The Measurement Model indicated close approximate fit (Model 2a, Table 1). Comparison of the No Item Bias Model with the

Measurement Model indicated the presence of gender-related item bias of the HADS-D. Step-by-step modification of the No Item Bias Model yielded the Final Model in which all bias was taken into account (Model 2c, Table 1). For Item 4 both uniform bias ($CHISQ_{diff}(1) = 16.55$, $p < .001$), and nonuniform bias was detected ($CHISQ_{diff}(1) = 14.47$, $p < .001$). In addition, nonuniform bias was detected for Item 10 ($CHISQ_{diff}(1) = 18.85$, $p < .001$). The Final Model showed equivalent fit as compared to the Measurement Model (see Table 1).

Age-related item bias. The Measurement Model showed close approximate fit (Model 4a, Table 1), but comparison with the No Item Bias Model indicated the presence of age-related item bias of the HADS-D (see Table 1). Uniform bias was detected in four items, and nonuniform bias was detected in three items, where one item showed both uniform and nonuniform bias. The Final Model, that included all apparent bias, showed close approximate fit (Model 4c). Although the Final Model did not yield equivalent fit as compared to the Measurement Model (see Table 1), freeing additional parameters did not significantly improve model fit. Uniform bias was detected for Item 4 ($CHISQ_{diff}(1) = 63.68$, $p < .001$), Item 6 $CHISQ_{diff}(1) = 102.97$, $p < .001$), Item 10 ($CHISQ_{diff}(1) = 40.12$, $p < .001$), and Item 14 ($CHISQ_{diff}(1) = 30.57$, $p < .001$). Nonuniform bias was detected for Item 4 ($CHISQ_{diff}(1) = 16.06$, $p < .001$), and Item 12 ($CHISQ_{diff}(1) = 20.51$, $p < .001$).

True differences between the groups. There were no significant differences between men and women ($d = 0.03$, $p = .64$) or between the age groups ($d = 0.03$, $p = .70$) with respect to their scores on the underlying Depression factor.

Multidimensional SEM approach

Results of Stage 1 indicated that the hypothesis of bivariate normality under equal thresholds was tenable for all combinations of items and exogenous variables. The estimated (polychoric) correlations, variances and means of all variables were used for subsequent analyses in Stage 2. In Stage 2, the Measurement Model that included both HADS subscales showed reasonable approximate fit (Model 5a, Table 1). The No Item Bias Model that included the variables Age and Gender did not show acceptable fit (Model 5b), indicating the presence of item bias (see Table 1). Uniform bias was detected in four items of the HADS-A, and six items of the HADS-D. The Final Model, that included all apparent bias, showed reasonable approximate fit (Model 5c, Table 1).

The anxiety subscale. Gender-related bias of the HADS-A was detected for Item 9 ($CHISQ_{diff}(1) = 24.2$, $p < .001$), and Item 11 ($CHISQ_{diff}(1) = 97.9$, $p < .001$). Age-related bias of the HADS-A was detected for Item 1 ($CHISQ_{diff}(1) = 64.0$, $p < .001$), and Item 13 ($CHISQ_{diff}(1) = 104.8$, $p < .001$).

The depression subscale. Gender-related bias of the HADS-D was detected for Item 2 ($CHISQ_{diff}(1) = 22.9, p < .001$), Item 6 ($CHISQ_{diff}(1) = 28.2, p < .001$) and Item 14 ($CHISQ_{diff}(1) = 28.9, p < .001$). Age-related bias of the HADS-D was detected for Item 4 ($CHISQ_{diff}(1) = 25.9, p < .001$), Item 6 ($CHISQ_{diff}(1) = 66.4, p < .001$), Item 8 ($CHISQ_{diff}(1) = 20.8, p < .001$), Item 10 ($CHISQ_{diff}(1) = 37.6, p < .001$), and Item 14 ($CHISQ_{diff}(1) = 52.5, p < .001$).

True differences and associations. Inspection of parameter estimates of The Final Model showed that there was a significant positive association between Anxiety and Depression ($r = 0.83$, $p < .001$), indicating that symptom severity with respect to Anxiety goes together with symptom severity with respect to Depression. There was a significant negative association between Age and Anxiety ($r = -0.24$, $p < .001$), indicating that older patients scored lower on Anxiety than younger patients. There was also a significant negative association between Gender and Anxiety ($r = -0.16$, $p < .001$), indicating that men scored lower on Anxiety than women. The association between Gender and Depression was negative, and between Age and Depression was positive, but not significant ($r = -0.04, p = .19$, and $r = 0.01, p = .74$ respectively). Lastly, there was a significant positive association between Age and Gender ($r = 0.11, p < .001$), indicating that men were - on average - significantly older than women.

Comparison with Cameron et al

With regards to the anxiety subscale of the HADS, both the multigroup SEM approach and multidimensional SEM approach identified uniform gender-related bias in Items 9 and 11, and uniform age-related bias in Items 1 and 13. In addition, with the multigroup SEM approach nonuniform age-related bias was detected in Item 3. These results are largely consistent with the results from Cameron et al., both in size and direction of detected bias. All methods identified age-related bias in Item 1, and the gender-related bias in Items 9 and 11 was also identified by the contingency tables method and the ordinal logistic regression method. The Rasch method detected uniform gender-related bias in Item 11, but the result was in the opposite direction. The (small) uniform age-related bias of Item 13 was not detected by the methods of Cameron et al., although the results of the contingency tables method and ordinal logistic regression method almost reached statistical significance for this item.

With regards to the depression subscale, the methods reported by Cameron et al. did not detect gender-related item bias, whereas both SEM methods did detect gender-related bias, but not in the same items. All three methods reported by Cameron et al. detected age-related bias in Items 6, 8, and 10. The results of the multigroup SEM approach confirmed the age-related bias in Items 6 and 10, and the results from the multidimensional SEM approach confirmed the age-related bias in all three items. However, the SEM approaches identified additional items with uniform and nonuniform age-related bias. Results of uniform age-related item bias from the multigroup SEM approach and multidimensional SEM approach are largely consistent, both in size and direction. Using the multigroup SEM approach we also detected nonuniform item bias in two items.

Table 1 | Goodness of overall model fit and difference in model fit of the models for gender- and age-related item bias detection models in Stage 2; for both the multigroup structural equation modeling approach, and the multidimensional structural equation modeling approach

Model		Df	CHISQ	p-value	RMSEA [90% CI]	Compared to	Df_{diff}	$CHISQ_{diff}$	p-value
Multigroup gender-related item bias detection									
Anxiety subscale									
1a	Measurement Model	28	50.64	.005	0.039 [0.021 ; 0.056]				
1b	No Item Bias Model	40	126.4	< .001	0.064 [0.051 ; 0.076]	Model 1a	12	75.76	< .001
1c	Final Model	38	81.29	< .001	0.046 [0.032 ; 0.060]	Model 1a	10	30.65	< .001
Depression subscale									
2a	Measurement Model	28	46.26	.016	0.035 [0.015 ; 0.052]				
2b	No Item Bias Model	40	120.9	< .001	0.062 [0.049 ; 0.074]	Model 2a	12	74.63	< .001
2c	Final Model	37	70.02	< .001	0.041 [0.026 ; 0.055]	Model 2a	9	23.76	.005
Multigroup age-related item bias detection									
Anxiety subscale									
3a	Measurement Model	28	61.02	< .001	0.047 [0.031 ; 0.063]				
3b	No Item Bias Model	40	163.2	< .001	0.076 [0.064 ; 0.088]	Model 3a	12	102.1	< .001

3c	Final Model	37	81.71	< .001	0.048 [0.04 ; 0.062]	Model 3a	9	20.69	.014
Depression subscale									
4a	Measurement Model	28	42.59	.038	0.031 [0.008 ; 0.049]				
4b	No Item Bias Model	40	357.2	< .001	0.122 [0.111 ; 0.134]	Model 4a	12	314.6	< .001
4c	Final Model	34	83.24	< .001	0.052 [0.038 ; 0.066]	Model 4a	6	40.65	< .001
Multidimensional gender- and age-related item bias detection									
Anxiety and depression subscale									
5a	Measurement Model	76	485.05		0.071 [0.065 ; 0.077]				
5b	No Item Bias Model	100	1029.8		0.093 [0.088 ; 0.098]				
5c	Final Model	88	455.71		0.063 [0.057 ; 0.068]				

Notes: $N = 1068$; Overall model fit and difference in fit was evaluated using WLS chi-square values that are provided in the standard LISREL output (denoted C2_NNT)

Table 2 | Results of gender- and age-related item bias detection in the anxiety and depression scales of the HADS questionnaire using the multigroup structural equation modeling (SEM-MG) and multidimensional structural equation modeling (SEM-MD) approaches. Results are compared to the item bias detection results as reported by Cameron et al. (2014) from the ordinal logistic regression method (LOGR), the Rasch model method (RASCH), and the contingency table method (CONT)

	Gender-related item bias					**Age-related item bias**				
Item	LOGR[1]	RASCH[2]	CONT[3]	SEM-MG[4]	SEM-MD[5]	LOGR[1]	RASCH[2]	CONT[3]	SEM-MG[4]	SEM-MD[5]
HADS-A										
1. I feel tense or wound up	-0.16	0.16	-1.03	-	-	**-0.77**	**-0.61**	**-3.78**	**-0.27**[a]	*-0.09*[a]
3. I get a ... feeling as if something awful...	-0.22	0.13	-1.21	-	-	0.08	0.05	0.54	*0.09*[b]	-
5. Worrying thoughts go through my mind	-0.10	0.09	-0.74	-	-	-0.16	-0.17	-1.01	-	-
7. I can sit at ease and feel relaxed	0.20	-0.20	1.68	-	-	0.06	0.06	-0.76	-	-
9. I get a .. feeling like 'butterflies' in the stomach	-0.49	0.44	**-3.64**	*-0.16*[a]	**-0.13**[a]	-0.36	-0.28	-2.12	-	-
11. I feel restless as if I have to be on the move	*0.58*	**-0.62**	**4.70**	**0.26**[a]	**0.12**[a]	0.25	0.40	1.99	-	-
13. I get sudden feelings of panic	-0.10	0.08	-0.79	-	-	0.58	0.34	3.24	*0.18*[a]	*0.07*[a]
HADS-D										
2. I still enjoy the things I used to enjoy	0.16	-0.07	0.72	-	*0.07*[a]	0.49	0.37	3.13	-	-
4. I can laugh and see the funny side of things	-0.06	0.07	-0.68	**-0.20**[a] *0.01*[b]	-	-0.33	-0.24	-1.11	**-0.66**[a] *0.01*[b]	**-0.13**[a]
6. I feel cheerful	0.35	-0.19	1.65	-	**0.11**[a]	**-1.11**	**-0.77**	**-5.16**	**-0.56**[a]	**-0.23**[a]
8. I feel as if I am slowed down	-0.17	0.16	-1.39	-	-	**0.92**	**1.03**	**6.72**	-	**0.14**[a]
10. I have lost interest in my appearance	-0.08	0.11	-0.67	*-0.01*[b]	-	*-0.60*	**-0.52**	**-3.66**	**-0.34**[a]	**-0.14**[a]
12. I look forward with enjoyment to things	-0.27	0.22	-2.34	-	-	0.39	0.18	2.00	*-0.01*[b]	-
14. I can enjoy .. book or radio or TV	-0.44	-0.35	2.91	-	**0.12**[a]	-0.56	-0.41	-2.46	**-0.24**[a]	**-0.18**[a]

Notes: [1] Log odds ratios are presented, where items were regarded as having important bias if the absolute magnitude of the log odds ratio was greater than 0.64 and $p < 0.001$. [2] Contrasts with absolute values greater than 0.50 and $p < 0.05$ were taken as an indication of important item bias. [3] Standardised Liu-Agresti Cumulative Common Log-Odds Ratios (LOR Z) are presented, where absolute values greater than 2 and $p < .001$ are considered important item bias. [4] Effect-sizes indices *d* are presented. For uniform item bias, these refer to the difference in intercept parameters between the groups, divided by the pooled standard deviation. For nonuniform item bias these refer to the difference in factor loading parameter multiplied with the difference in common factor means between the groups, divided by the pooled standard deviation. Effect-sizes larger than .20 and $p < .001$ are indicative of important item bias. [5] Effect-sizes indices *r* are presented, which are the standardized direct effect of Gender/Age on the specific item. Effect-sizes larger than .10 and $p < .001$ are indicative of important item bias. [a] Uniform item bias. [b] Nonuniform item bias. Results meeting the criteria for important item bias are marked in bold, results meeting only the significance criterion are marked in italics.

Discussion

Two different SEM approaches were applied to detect gender- and age item bias in the anxiety and depression subscales of the HADS, to account for item bias, and to more validly compare patients' anxiety and depression. Results from both the multigroup SEM approach and the multidimensional SEM approach confirmed the results of the ordinal logistic regression, item response theory (IRT), and contingency tables methods reported by Cameron et al. (2014). However, in general, the SEM approaches identified more items. These differences may indicate that the SEM method has more power to detect effects. Below we provide substantive interpretation of item bias that was consistently detected by the different approaches.

With regards to the anxiety subscale of the HADS, the detected gender-bias indicates that men, compared to women, scored relatively low on Item 9, "I get a ... feeling like 'butterflies' in the stomach", and relatively high on Item 11, "I feel restless as if I have to be on the move". This result indicates that anxiety symptoms manifested themselves differently in men as compared to women, where restlessness was more prevalent in men and 'butterflies' in the stomach were more prevalent in women, relative to the anxiety level. In addition, we found that patients aged above 65, as compared to patients aged below 65, scored relatively high on Item 1, "I feel tense or wound up", and Item 13, "I get sudden feelings of panic". These results indicate that older patients experienced more symptoms of panic and tenseness, relative to their level of anxiety.

With regards to the depression subscale of the HADS, age-related bias was detected for Items 6, 8 and 10. Taking into account reversed scoring of contra-indicative items, we found that patients aged above 65, as compared to patients aged below 65, scored relatively high on Item 6 ("I feel cheerful"), Item 8 ("I feel as if I am slowed down"), and relatively low on Item 10 ("I have lost interest in my appearance"). These results show that older patients more easily indicated to be cheerful, but also that they were slowed down, whereas they indicated losing less interest in their appearance, relative to their level of depression. In addition, we also found that patients aged above 65, as compared to patients aged below 65, scored relatively high on Item 4 ("I can laugh and see the funny side of things"), and Item 14 ("I can enjoy ... book or radio or TV"). These results show that older patients more easily indicated to see the funny side of things and enjoy a book, relative to their level of depression.

The results of Cameron et al. (2014) were generally consistent with the results of both SEM approaches applied in the present paper. To further investigate and compare the appropriateness of the different item bias detection methods, a simulation study would be required to investigate whether uniform and nonuniform item bias can be correctly identified. In such a simulation study one could, for example, investigate the performance of these different approaches under different circumstances, e.g., the size of the item bias, the direction of the item bias, the type of item bias, the number of items affected by bias, etc.

The multigroup SEM approach and the multidimensional SEM approach showed some differences with respect to the detection of item bias. Differences between the two SEM

approaches can occur because of several reasons. One reason may be that when age and gender are related, then age-related item bias may be detected in the multigroup SEM approach only because there exists gender-related item bias (or vice versa), whereas in the multidimensional SEM approach possible relations between gender and age are taken into account. In the present application, the age-related item bias that was detected for Item 4 and Item 10 using the multidimensional SEM approach may have sufficiently explained the apparent gender differences on these items that were detected using the multigroup SEM approach. For two of the three additional items that were identified with uniform gender bias in the multidimensional SEM approach, results were in the opposite direction of the age-related bias that was detected for these items. These effects might have been obscured in the multigroup SEM approach due to the association between gender and age.

Another difference between the multigroup SEM approach and the multidimensional SEM approach that was applied, is that the latter did not include the investigation of nonuniform item bias. Although it has been shown that investigation of nonuniform item bias is possible by including interaction-terms between the underlying trait of interest and the other exogenous variables, these type of extensions are not easily implemented and were therefore not applied in the present paper. Even though possible nonuniform item bias thus remained undetected, it did not seem to have influenced the identification of uniform item bias.

Finally, another reason for possible differences between the two SEM approaches is that the multidimensional SEM approach may have larger power to detect uniform item bias. Further investigation and comparison of the multigroup and multidimensional models is needed to substantiate the appropriateness of these different methods under different circumstances, and better explain possible differences in the detected effects.

To conclude, both the multigroup SEM approach and multidimensional SEM approach can be applied to detect bias in observed item scores. Advantages of the multigroup SEM approach over other item bias detection methods (e.g., the ordinal logistic regression, IRT, and contingency tables methods used by Cameron et al.) are that it uses a latent trait operationalization, it can detect both uniform and nonuniform bias, and possible item bias can be taken into account to assess true differences between groups. In addition, the extension to multidimensional models enables the investigation of item bias with respect to any factor or variable (e.g., continuous or categorical, latent or manifest), where continuous variables can be included in the model without categorizing them, and item bias can be investigated with respect to several variables simultaneously. Therefore, the SEM method provides a flexible tool for the investigation of item bias in health-related questionnaires, and may thus ultimately provide a more valid comparison of patients' well-being that is relevant for both research and clinical practice.

Acknowledgements

We would like to thank I. M. Cameron from the University of Aberdeen for making the data available for secondary analysis.