Genetic determinants of eating disorders

Slof-Op 't Landt, M.C.T.

Chapter 3

Sex differences in sum score may be hard to interpret: the importance of measurement invariance

This chapter was previously published:
Slof-Op 't Landt MC, Dolan, C. V., Rebollo-Mesa, I., Bartels, M., van Furth, E. F., Van Beijsterveldt, C. E., Meulenbelt, I., Slagboom, P. E., & Boomsma, D. I. (2009). *Assessment, 16,* 415-423.

Abstract

In most assessment instruments, distinct items are designed to measure a trait, and the sum score of these items serves as an approximation of an individual's trait score. In interpreting group differences with respect to sum scores, the instrument should measure the same underlying trait across groups (e.g., male/female, young/old). Differences with respect to the sum score should accurately reflect differences in the latent trait of interest. A necessary condition for this is that the instrument is *measurement invariant*. In the current study we illustrated a stepwise approach for testing measurement invariance with respect to sex in a 4-item instrument designed to assess disordered eating behavior (DEB-scale) in a large epidemiological sample (1195 men and 1507 women). Our approach can be applied to other phenotypes for which group differences are expected. Any analysis of such variables may be subject to measurement bias if a lack of measurement invariance between grouping variables goes undetected.

Questionnaires are often used to assess psychological and behavioral traits on a quantitative scale. Well-known examples are the Beck Depression Inventory (Beck et al., 1961), Eysenck EPQ scales (Eysenck & Eysenck, 1975) and, the Temperament and Character Inventory (Cloninger et al., 1993). In these assessment instruments, items are designed to measure an underlying trait or latent (i.e., unobserved) variable and scores on the items are summed to derive a total score on the trait of interest. The Diagnostic and Statistical Manual of Mental Disorders (4th ed., American Psychiatric Association, 1994) also employs a weighted sum score in diagnosing psychiatric disorders.

When comparing groups, it is vital that an instrument measures the same underlying trait across groups (e.g., male/female, young/old). Observed group differences in the sum scores should accurately reflect group differences with respect to the latent variable. A necessary condition for this is that the instrument displays measurement invariance with respect to the groups under consideration (Mellenbergh, 1989; Meredith, 1993). If there is a sex difference with respect to the latent trait, men should for example score lower on all the items of the instrument measuring this trait. If however, men score lower on all the items but one, this one item displays differential item functioning, and the scale is not measurement invariant with respect to sex (Dolan, 2000; Mellenbergh, 1989; Meredith, 1993; Millsap & Yun-Tein, 2004). In that case, group differences in sum scores reflect, at least in part, measurement bias. The interpretation of differences between groups with respect to the sum scores thus hinges on the establishment of measurement invariance, or at least on the understanding of the violations, if any, of measurement invariance. Ideally, differences in sum scores should reflect true differences in the latent variable that the psychometric instrument purports to measure.

Measurement invariance can be investigated by fitting a measurement model that relates item scores to the underlying trait(s) across groups. Several methods have been suggested for both continuous and categorical variables (Dolan, 2000; Mellenbergh, 1989; Meredith, 1993; Millsap & Yun-Tein, 2004; Muthen & Asparouhov, 2002; Muthén & Muthén, 2005). In the current study we described a stepwise approach that was derived from previous studies to investigate measurement invariance for ordered categorical items. Our goal was to provide a comprehensive overview of the different steps accumulating into a model of complete measurement invariance. To illustrate this approach, we investigated whether a four item instrument, designed to measure disordered eating behavior is measurement invariant with respect to sex. As eating disorders mainly affect young women (90 – 95% of cases) (Fairburn & Harrison, 2003; Hoek, 1993; Van Hoeken et al., 1998), one might expect sex differences in the endorsement of the four eating disorder items.

Multi-group discrete factor analyses were applied to test whether the disordered eating behavior instrument is measurement invariant with respect to sex.

Method

Participants

All participants were registered with the Netherlands Twin Registry, which is maintained at the Department of Biological Psychology at the VU University in Amsterdam (Bartels et al., 2007; Boomsma et al., 2006). In this study, we used data from the 1986-1992 birth cohorts. In January 2005, questionnaires were sent to adolescent twins (mean age 15.2, *SD*=1.3) and their non-twin siblings (mean age 16.7, *SD*=2.8). The twins and siblings were asked to complete a survey containing items relevant for eating disorders. Questionnaires were sent to 2000 families. A total of 2175 twins (twin response rate 54.4%) and 527 siblings from 1144 families returned the questionnaire (family response rate 57.2 %). The total sample consisted of 1195 men and 1507 women (956 male twins, 1219 female twins, 239 brothers and 288 sisters, respectively), mean age was 15.5 (*SD*=1.8).

Measures

Participants filled out a self-report questionnaire containing measures of health and behavior (Bartels et al., 2007; Boomsma et al., 2006). The eating disorder section included four items: 1) dieting (Q: Have you ever gone on a diet to lose weight or to stop gaining weight?); 2) fear of weight gain (Q: How afraid are you to gain weight or become fat?); 3) importance of body weight or shape on self-evaluation (Q: How important are body weight and/or shape in how you feel about yourself?); 4) binge eating (Q: Have you ever had episodes of binge eating?). Responses were given on five point Likert-scales, ranging from ‘never’ to ‘always’ for dieting (DIET), from ‘not afraid’ to ‘extremely afraid’ for fear of weight gain (FEAR), from ‘not important’ to ‘most important’ for importance of body weight and shape on self-evaluation (ISE), and from ‘never’ to ‘more than once a week’ for binge eating (BE). For the multi-group confirmatory factor analyses it was essential that, for every item, each category was endorsed by both groups. Because none of the men reported that they were always on a diet, the fourth and fifth categories of the dieting item were merged. As a consequence, three items with five categories and one item with four categories were used in the analyses.

Data Analysis

We performed multi-group confirmatory factor analyses to establish whether the four eating disorder items formed a uni-dimensional scale, and whether the scale was measurement invariant with respect to sex. To conduct a confirmatory factor analysis, a minimum of three items is required. Measurement invariance with respect to sex held if the probability of a certain response on a given item was the same for all participants with the same value on the underlying trait (disordered eating behavior [DEB]) regardless of the sex of the participant. This definition gave rise to a highly constrained multi-group factor model (Chen et al., 2005; Meredith, 1993; Millsap & Yun-Tein, 2004). To establish measurement invariance, we fitted several increasingly restrictive models derived from approaches described in previous studies (Dolan, 2000; Mellenbergh, 1989; Meredith, 1993; Millsap & Yun-Tein, 2004; Muthen & Asparouhov, 2002; Muthén & Muthén, 2005), cumulating in this highly constrained model.

In the first step, a saturated model was fitted to the data simply to obtain estimates of the item thresholds and the polychoric correlation among items. To this end, we assumed that a latent continuous variable, called the liability, was underlying the responses to each discrete item. Assuming the liability underlying each item was standard normally distributed, the discrete responses were modeled to items by estimating thresholds on the standard normal distributions of the liability (3 thresholds for the DIET item, and 4 thresholds for the other three items). The positions of these thresholds determined the marginal response probabilities of each item. In addition, the (polychoric) correlations among the liability underlying the four items were estimated. Thresholds and correlations were estimated separately in men and women.

In the second model it was tested whether the four items were uni-dimensional in men and women. The four continuous latent liabilities were regressed on a single common factor, without imposing any equality constraints over sex. Thresholds in men were constrained to equal those in the women. By imposing this constraint, the thresholds were estimated on a common metric. The distribution of the liability for each item was standard normal in the women as in model 1. In the men the means and variances of the liability underlying the four items were estimated freely. Thus, in this step we fitted a single factor model to the correlation matrix of the liabilities in the women, and a single factor model to the covariance matrix of the liabilities in the men. In both sexes, the common factor was scaled to have a mean of zero and a variance of one (i.e., standard scaling constraints in the common factor model). By estimating all the factor loadings freely, the item reliability in the women and the men were obtained separately. Note that these reliability estimates need not be equal over sex.

In model 3, the factor loadings were constrained to be equal over sex. This constraint allowed estimation of the variance of the common factor in one group (men), while retaining the scaling constraint (variance of factor equal to one) in the other group (women). We thus allowed for a difference in common factor variance between men and women. This model included sex differences in the residual variances of the items, in the liability means and in the common factor variance.

In model 4 mean liabilities (intercepts) in the male sample were constrained at zero, and the common factor mean was estimated. As before, the mean liabilities and common factor mean were fixed to zero in women. In the preceding model the estimated mean in liabilities in men gave an indication of the sex differences per item. By fixing these intercepts at zero in men, while freely estimating the mean of the common factor, any sex difference in means of the liabilities was explained by a difference in the mean of the common factor, i.e. a difference with respect to the latent variable of interest.

In model 5, we added the final constraint of 'invariance of residual variances over sex'. As a consequence, the amount of the variance in the separate items that was not explained by the common factor was constrained to be equal in the women and men. This model represented full measurement invariance. Note that in this model any observed sex difference in the observed test scores was attributable to a difference with respect to the latent variable that we purported to measure. With respect to the interpretation of sex differences in test scores, model 5 represented the ideal. Model 4 represented a weaker form of invariance in which sex differences in the residuals were permitted. Model 4 was still useful as it allowed us to interpret sex differences in the mean scale score as a manifestation of a mean difference with respect to the latent variable. Weaker forms of measurement invariance are entertained in the literature (e.g., model 3: equality of factor loadings), but we did not consider these to be sufficient for the interpretation of sex differences with respect to the test scores (Meredith, 1993).

All analyses were performed in Mplus 4.0 (Muthen & Asparouhov, 2002; Muthén & Muthén, 2005). Because our sample consisted of families, the individual cases were not independent. To correct for the effect of this dependence on the standard errors and overall goodness of fit indices, we used the Weighted Least Square with mean adjusted Chi-square test statistics (WLSM) in combination with the 'Complex' option in Mplus. The latter corrects the statistical effect of clustering on the results. Rebollo et al. (2006) found this method to be satisfactory to correct for dependency due to family grouping.

As suggested by Schermelleh-Engel, Moosbrugger and Müller (2003), several fit statistics were used to evaluate the fit of the models; hierarchical Chi-square tests, the comparative fit index (CFI), and the root mean square error of approximation (RMSEA).

For the hierarchical Chi-square test, the difference between the Chi-square test statistics obtained for each model yielded a new Chi-square value with degrees of freedom equal to the difference in the number of parameters in the two models. In the WLSM approach in Mplus, the reported Chi-squares were mean adjusted and a scaling correction factor was applied for each model. As a consequence, in calculating the Chi-square difference test, scaling correction factors had to be entered into the equation (Asparouhov & Muthen, 2006). According to the principle of parsimony, models with fewer parameters are preferred, if they do not give a significant deterioration of the fit. Significance can be determined on statistical grounds, but in structural equation modeling, rules of thumb are usually used (Schermelleh-Engel et al., 2003). The CFI ranges from zero to one with higher values indicating better fit; for a good model fit the CFI should be above 0.97, and values greater than 0.95 indicate an acceptable fit (Schermelleh-Engel et al., 2003). The RMSEA is a measure of closeness of fit, and provides a measure of discrepancy per degree of freedom. A value of 0.05 or smaller indicates a close fit, and values between 0.05 and 0.08 indicate an acceptable fit (Jöreskog, 1993; Schermelleh-Engel et al., 2003).

There were 257 persons (n=127 men and n=130 women) who completed the survey twice with an interval of six months. Retest data obtained in this group will serve to estimate stability of the test scores. The reliability of the eating disorder items was estimated separately in men and women. Polychoric correlations between the two occasions of measurement were calculated for each item using Mplus.

Results

To evaluate how often the different eating disorder attitudes and behaviors were endorsed, we calculated the frequencies of the item scores greater than three in the adolescent twins and their non-twin siblings for the four items. These frequencies showed significant sex differences for three features (p<0.001). For the DIET item, 0.4% of the men compared to 3.4% of the women had been on a diet often or always. Few men (1.3%) reported being very or extremely afraid to gain weight or become fat (FEAR). In women this item was endorsed more often with 8.7%. A large proportion of both men and women reported that “their body weight and or shape played an important role in how they felt about themselves” (ISE). The frequency of this feature was 40.9% in the women compared to 26.8% in the men. No sex differences were found for the BE item, 5.1% of the women and 5.5% of the men reported having binge eating episodes at least once a week.

In model 1 polychoric correlations among items, and the thresholds for each item were estimated per sex. These are reported in Table 3.1. Small to moderate correlations between

the items were found in both sexes. Although the magnitude of the correlations differed between groups, similar patterns were observed with the highest correlation between DIET and FEAR and the lowest between ISE and BE. The thresholds of the liabilities represent the cut-points of the response categories in the corresponding ordinal items on a sex-specific z-scale. The mainly positive thresholds indicate that the majority of women and men did not engage in eating disordered behaviors and/or attitudes.

Table 3.1 Correlations and thresholds for women and men (saturated model).

	DIET[b]	FEAR[c]	ISE[d]	BE[e]
Correlations[a]				
DIET[b]	1.00	0.59 (0.52,0.66)	0.39 (0.30,0.48)	0.41 (0.31,0.51)
FEAR[c]	0.53 (0.38,0.67)	1.00	0.59 (0.54,0.64)	0.33 (0.24,0.41)
ISE[d]	0.27 (0.13,0.40)	0.39 (0.29,0.48)	1.00	0.27 (0.19,0.36)
BE[e]	0.22 (0.03,0.41)	0.20 (0.06,0.34)	0.16 (0.06,0.27)	1.00
Women				
Threshold 1	0.68 (0.57,0.78)	-0.43 (-0.52,-0.33)	-1.54 (-1.68,-1.40)	0.64 (0.54,0.74)
Threshold 2	1.36 (1.24,1.49)	0.67 (0.57,0.77)	-0.56 (-0.66,-0.46)	1.12 (1.00,1.23)
Threshold 3	1.83 (1.66,1.99)	1.36 (1.24,1.49)	0.23 (0.14,0.32)	1.63 (1.48,1.78)
Threshold 4	-	2.13 (1.93,2.33)	1.88 (1.70,2.06)	2.08 (1.87,2.29)
Men				
Threshold 1	1.62 (1.44,1.80)	0.72 (0.60,0.83)	-0.94 (-1.06,-0.83)	0.95 (0.83,1.07)
Threshold 2	2.29 (2.02,2.56)	1.71 (1.52,1.89)	-0.10 (-0.20,0.001)	1.27 (1.13,1.40)
Threshold 3	2.64 (2.25,3.02)	2.24 (1.97,2.51)	0.58 (0.48,0.69)	1.60 (1.44,1.76)
Threshold 4	-	2.71 (2.28,3.14)	1.94 (1.73,2.15)	1.87 (1.68,2.06)

[a] The correlations in the women are listed above the diagonal, the correlations in the men are listed below the diagonal. Numbers in parentheses represent 95% confidence intervals.
The thresholds are estimated on a sex-specific z-scale.
[b] DIET: Dieting
[c] FEAR: Fear of weight gain
[d] ISE: Importance of body weight or shape in self-evaluation
[e] BE: Binge eating

In Table 3.2, fit statistics of the nested models are given. Model 2, which tested whether one factor could account for the correlations among the four eating disorder variables, fitted significantly worse compared to model 1 according to the chi-square. However, both the RMSEA and the CFI indicated a good fit of this model. The parameter estimates of model 2 are presented in Table 3.3. The factor loadings of DIET and BE were comparable between men and women. On the other hand, the factor loading in the men for FEAR was higher and for ISE was lower compared to the women. The least reliable item was BE, while the FEAR item had the highest reliability.

Table 3.2 Model fit statistics

Model	χ2	df	CFI [a]	RMSEA [b]	CM [c]	Δχ2 [d]	Δdf [e]	*p*
Model 1 (saturated)	0.00	0	1.00	0.00	-	-	-	-
Model 2 (one factor model)	37.98	11	0.99	0.04	1	37.98	11	0.0001
Model 3	35.62	14	0.99	0.03	2	2.32	3	0.51
Model 4	101.07	17	0.96	0.06	3	50.99	3	0.0001
Model 5 (full measurement invariance)	246.53	21	0.90	0.09	4	99.57	4	0.0001

[a] CFI: Comparative Fit Index

[b] RMSEA: Root Mean Square of Error of Approximation

[c] CM: Compared to model

[d] Δχ2: Chi-square test statistic between two models adjusted for scaling correction factor

[e] Δdf: degrees of freedom for the Chi-square difference test

The estimates of the mean liability in men were all significantly lower than zero. As these means were fixed to zero in the women, we established, as expected, that the men scored lower than the women on all eating disorder items. The estimated variances of the liability of FEAR, ISE, and BE were significantly smaller than one in the men. The variances were fixed at one in the women.

Table 3.3 Parameter estimates for model 2 in the female reference group and the male group.

	DIET[a]	FEAR[b]	ISE[c]	BE[d]
Women				
Factor loading	0.68 (0.60, 0.75)	0.88 (0.81, 0.94)	0.66 (0.59, 0.72)	0.44 (0.35, 0.52)
Mean	0	0	0	0
Variance	1	1	1	1
Reliability	0.46	0.77	0.43	0.19
Men				
Factor loading	0.69 (0.35, 1.03)	0.97 (0.71, 1.24)	0.55 (0.41, 0.70)	0.45 (0.19, 0.71)
Mean	-1.11 (-1.83, -0.39)	-1.30 (-1.56, -1.05)	-0.44 (-0.57, -0.31)	-0.84 (-1.28, -0.40)
Variance	0.91 (0.59, 1.24)	0.84 (0.70, 0.98)	0.85 (0.77, 0.93)	0.65 (0.50, 0.79)
Reliability	0.48	0.94	0.30	0.20

Numbers in parentheses represent 95% confidence intervals for the factor loadings and residual variances

[a] DIET: Dieting

[b] FEAR: Fear of weight gain

[c] ISE: Importance of body weight or shape in self-evaluation

[d] BE: Binge eating

The Chi-square test statistic suggested some violation of uni-dimensionality (model 2). But because both the RMSEA and the CFI indicated a good fit, the invariance of factor

loadings across sexes was tested next. For this model, all three fit statistics indicated a good fit. The estimate of variance of the common factor (disordered eating behavior (DEB)) in the male group was 0.96. Given the 95% confidence interval (CI) of 0.62 and 1.30, we concluded that the variance was not significantly different between the men and women in model 3.

In model 4, the mean of the liabilities were constrained to be zero in men (as they were in women). The mean of the common factor was fixed to zero in the women, as before, and estimated freely in the men. This model did not fit very well in comparison to model 3. The Chi-square test statistic indicated a significantly worse fit for this model. However, the fit was acceptable according to the RMSEA and the CFI. The estimated common factor mean in the men was -0.99, which differed significantly from zero (95% CI -1.18 - -0.80). In other words, the mean of DEB was lower in men than in women (factor mean fixed at zero).

Because the fit of model 4 was acceptable based on the RMSEA and the CFI, the final model of complete measurement invariance was tested. In this fifth model, the residual variances were also constrained to be equal across the groups. The Chi-square statistic indicated deterioration in fit compared to model 4. In addition, the CFI and the RMSEA indicated a bad fit. This implied that the eating disorder items were not fully measurement invariant with respect to sex. The variances presented in Table 3.3, give an indication of which item might be underlying this bad fit. The variance of BE showed the largest deviation from 1, suggesting that the greatest difference between both groups in residual variance was observed for this item.

Finally the stability of the item responses and the DEB total score were considered. The four eating disorder items were moderately to highly correlated over a period of six months. The polychoric correlation was 0.59 (95% CI 0.28-0.89) for DIET, 0.75 (95% CI 0.59-0.90) for FEAR, 0.56 (95% CI 0.41-0.71) for ISE, and 0.74 (95% CI 0.55-0.93) for BE in men. In women, the polychoric correlation was 0.75 (95% CI 0.60-0.89) for DIET, 0.67 (95% CI 0.55-0.79) for FEAR, 0.43 (95% CI 0.27-0.59) for ISE, and 0.58 (95% CI 0.42-0.74) for BE.

Discussion

In most assessment instruments, distinct items are designed to measure a trait, and the sum score of these items serves as an approximation of an individual’s trait score. The interpretation of differences between groups with respect to these sum scores hinges on the establishment of measurement invariance. Ideally, differences in sum scores should reflect

true differences in the latent variable that the psychometric instrument purports to measure. If there is a lack of measurement invariance, group differences in sum scores reflect, at least in part, measurement bias.

We described a stepwise multi-group confirmatory factor analysis to investigate measurement invariance for categorical items with respect to a grouping variable. Previously, several methods have been reported to test for measurement invariance both for continuous and categorical items (Dolan, 2000; Mellenbergh, 1989; Meredith, 1993; Millsap & Yun-Tein, 2004; Muthen & Asparouhov, 2002; Muthén & Muthén, 2005). All these methods cumulated in an identical highly constrained model in which strict factorial invariance, or complete measurement invariance, was tested. However, the number and order of the constraints in the intermediate models differed between the reported methods. In contrast to previous studies, our analysis began by fitting a saturated model to the data, to obtain estimates of the polychoric correlation among items and the thresholds for each item. The second model, which tested for uni-dimensionality of the items, was more comparable to the baseline models described by other groups (Millsap & Yun-Tein, 2004; Muthen & Asparouhov, 2002; Muthén & Muthén, 2005), although there was a difference in the constraints. In our model, thresholds were constrained across groups, while factor loadings were estimated freely. This enabled us to calculate the reliability of the separate item scores. Means and variances of the liabilities provided insight in the between-group differences. In the third model, both item thresholds and factor loadings were constrained to be equal across groups. The between-group differences in this model were represented by the residual variances of the items, the liability means and by the common factor variance. In addition to the previous constraints, the liability means were constrained at zero in all groups in model 4. Within this model, any group difference in the means of the latent indicators would be explained by a difference in the mean of the common factor. This model represented a weaker form of invariance in which group differences were permitted in the residuals, and was similar to the third model described by Millsap et al. (2004). The final model of strict factorial invariance, added the constraint of invariance of residual variances over groups; i.e. the amount of the variance in each item that was not explained by the common factor was constrained to be equal in the groups.

The method was illustrated by investigating whether a scale comprised of four eating disorder items was measurement invariant with respect to sex. The model of full measurement invariance with respect to sex (model 5), did not fit the data well. If this model had fitted, the probability of a certain response on a given item would have been the same for all participants with the same value on the underlying trait (DEB) regardless of the sex of the participant. However, this was not the case. The underlying common factor might

not be the only source of difference between the sexes with respect to the four items. The sum score based on the four eating disorder items therefore cannot be taken to represent exactly the same underlying trait in men and women. This means that sex differences in this sum score might be due to measurement bias instead of a true difference in the underlying trait.

What implication does this finding have for existing eating disorder measurement instruments? We acknowledge that a scale consisting of four items might not be ideal to measure the underlying latent trait in eating disorders. However, in large epidemiological studies such as becoming common for gene finding, short scales might be a requirement to obtain phenotyping in sufficiently large samples. With the selection of the items we have tried to capture a variety of eating disorder symptoms. Three of the items (FEAR, ISE and BE) used in this study are based on eating disorder criteria from the Diagnostic and Statistical Manual of Mental Disorders (4^{th} ed., American Psychiatric Association, 1994). The fourth item (DIET) has been identified as a potent risk factor (Jacobi et al., 2004). However, one eating disorder symptom, compensatory behavior, is missing in our assessment instrument.

There has been a lot of debate about whether eating disorders are dimensional like proposed in the "continuum of eating disorders" (Fairburn & Harrison, 2003; Hay & Fairburn, 1998), or whether they are discrete syndromes (Williamson et al., 2005). Some studies suggest that eating disorders can be conceptualized as having at least two latent features (Williamson et al., 2002; Williamson et al., 2005); binge eating, and general psychopathology. Accordingly, the FEAR, DIET, and ISE items would load on one factor, and the BE item would load on a second factor. The correlations presented in Table 3.1, however, show substantial correlations between DIET and BE, especially in women (0.41). Bulimic behavior has been correlated with dieting and body concerns in several other studies (Williamson et al., 2005), although this correlation appears to exist exclusively in nonclinical samples. Since our sample is also nonclinical, this may be the cause of the high correlation between DIET and BE. Hence, the factor structure discussed above might not be suitable in nonclinical groups. On the other hand, the low reliability of the BE item and the fact that the variance of this item showed the largest deviation from one in model 2, might be supportive of the two factor structure underlying eating disorders. However, investigating partial measurement invariance by omitting the final constraints on the BE item did not lead to a model of strict factorial invariance for the remaining three items.

The finding of a lack of strict factorial invariance in the 4-item DEB scale might not generalize to existing eating disorder scales. However, this form of measurement invariance has never been tested in the eating disorder field. Many studies have used both exploratory

and confirmatory factor analysis to test whether existing measurement instruments have the same factor structure across, for example, different types of patients, and different ethnic groups, and to establish different factors within eating disorders (Calugi et al., 2006; Fernandez et al., 2006; Hrabosky et al., 2008; Lee et al., 2007; Peterson et al., 2007; Varnado et al., 1995; Wade et al., 2008a; Williamson et al., 2002; Williamson et al., 2005). Until now, only one study has investigated measurement equivalence (Warren et al., 2008). Warren et al. tested for the equivalence of factor loadings for the Body Shape Questionnaire in American and Spanish women with and without an eating disorder diagnosis. For a subscale of 10 items, the constraint of invariant factor loadings fitted the data well. However, because the intercepts were not constrained to equivalence in this study, the scores in the different groups may not have the same origin (Chen et al., 2005). Thus, differences on factor means between groups could still be caused by measurement bias.

The responses to the four eating disorder items were fairly stable over a six month period, with correlations ranging from 0.43 for ISE in the women to 0.75 for FEAR in the men and DIET in the women. The prevalence for the DIET, FEAR, and BE item were low to moderate. The prevalence of ISE was substantially higher. Comparable rates were found in other population-based studies in adolescents with the exception of the DIET item, which had a lower prevalence (Kjelsas et al., 2004; Neumark-Sztainer et al., 2007; Rowe et al., 2002; Silberg & Bulik, 2005). Because of the low endorsement rates of dieting in the men, we had to merge the fourth and fifth category for the DIET item. As a consequence, the number of response frequencies differed between the four items. This difference in response categories does not appear to impact the results. When all items are merged into four or even three categories, the same results were found throughout the different steps of the confirmatory factor analyses. Comparable correlations, thresholds and factor loadings for the four items were found. In addition, the model of weak measurement invariance (model 4) remained the best-fitting model.

The framework we presented in this paper can serve as a valuable tool for examining the psychometric qualities of other interviews and questionnaires with respect to sex. In addition, other kinds of grouping variables (e.g. age, level of education) can also be studied using this method. An advantage of our approach is that it provides a better understanding of the consequences of the different constraints per model. As a consequence, it gives a better insight into the violations of measurement invariance, and the underlying causes of this measurement bias. It is essential to test for measurement invariance before sum scores or scale scores are used to compare groups. This is not only the case in the eating disorder field, but applies to other fields of research as well.