Arch Public Health
2009, 67, 15-29

Measurement Equivalence of the CES-D 8 in the General Population in Belgium: a Gender Perspective

by

Van de Velde S[1], Levecque K[1,2], Bracke P[1]

Abstract

International research consistently finds gender differences in depression, but do women genuinely experience more complaints or are the findings contaminated by group-specific elements unrelated to depression but affecting its measurement? The study of gender differences in depression depends on the measurement quality of the instrument used to evaluate depression. In the present study we test the measurement equivalence of a shorter version of a commonly used instrument in mental health research, the Center for Epidemiologic Studies - Depression Scale (CES-D), using data from the Belgian sample of the third round of the European Social Survey (N=1794). Evidence for measurement invariance can be established within the multigroup confirmatory factor analysis framework. This method allows us to evaluate a nested hierarchy of hypotheses to test different levels of cross-group measurement invariance: configural, metric, scalar and residual invariance, and clarifies under what conditions meaningful comparisons between the male and female respondents can be made. The best fitting factor model is then used to estimate the 'true' prevalence of depressive symptoms for both groups. In our study measurement equivalence is established at all levels, indicating that the current depression scale allows defensible quantitative gender comparisons. Our data also confirm the epidemiological finding that women report more complaints of depression than men.

Keywords

Depression, factor analysis, gender, validation, psychometrics

Introduction

Gender differences in depression: a partly artificial fact?

According to the World Health Organization, depression is the most common mental health problem in the Western world (1). A recurrent finding in international literature is that there is

[1] Department of Sociology, Ghent University, Ghent, Belgium

[2] Research Fellow FWO Flanders

Correspondence: sarah.vandevelde@ugent.be

a 1.5 to 3 times higher prevalence of depression in women compared to men (2-5). This is true for inpatient and outpatient as well as general population studies. The pattern of a higher prevalence of depression in women compared to men has been consistent across nations, cultures and population groups, in studies using different methods and measurement instruments and for a diversity of incidence and prevalence indicators (3, 4).

In the Belgian context, the existence of gender differences in depression has been confirmed in patient samples (6), and in recent years in the general population. General population research was mostly based on the Mini International Neuropsychiatric Interview in the Belgian sample of the Depression Research in European Society-survey (7, 8), and on the International Diagnostic Interview on the Belgian data of the European Study of the Epidemiology of Mental Disorders (9, 10). In addition, gender differences were assessed by a depression-subscale of the Health and Daily Living Form in the Panel Study of Belgian Households (11-13), and the Symptom Checklist 90-Revised in the Belgian Health Interview Surveys of 2001 and 2004 (14, 15). Although this recurrent epidemiological finding indicates the existence of a true gender difference in the prevalence of depression in Belgium and abroad, the possibility remains that the observed difference between men and women is partly due to possible measurement variance. In the present study, we aim to evaluate the measurement invariance of a depression scale as a tool for making cross-gender comparisons. Our estimates of gender differences in depression in the general Belgian population will therefore reflect true differences between men and women, rather than being contaminated by possible group-specific attributes unrelated to depression.

In the present study, we made use of the Belgian sample of the third round of the European Social Survey (ESS 3) (16), organised in 2006 and 2007. Depression is assessed by the Center for Epidemiologic Studies - Depression Scale or CES-D (17). Previous studies on the measurement equivalence of the 20-item CES-D scale are ambivalent, with several studies pointing towards a gender bias (18-22) while other studies suggest measurement equivalence of the scale across gender groups (23, 24). Several other measurement inventories for depression also pointed towards a gender bias (5, 13, 25, 26). In the ESS 3, depression was not assessed using the CES-D 20, but respondents were administered an 8-item version of the CES-D. While the CES-D 20 is used extensively in international research, the CES-D 8 has seldom been used before.

Measurement equivalence and factorial invariance

When depression is measured using a multi-item self-report instrument such as the CES-D, each item is considered an imperfect measure of one of the symptoms of depression, but as a whole, the set of items is hoped to provide a valid indirect assessment of a latent construct called depression. When the items are summed to form a composite measure, it is also assumed that the total measurement score will be more reliable than single item scores (27).

Measurement equivalence or invariance is the condition that is attained when individuals with equivalent true scores on a measurement instrument for a latent construct have the same probability of a particular observed score on an associated test (28). Measurement equivalence or invariance is the broader concept that subsumes factorial invariance. The latter is a measurement equivalence approximation that can be tested with confirmatory factor analysis (CFA), a special case of structural equation modelling. CFA is an excellent way of determining whether a hypothesised common factor underlies a scale and is currently the methodology of choice for assessed factorial invariance (29). In the jargon of factor analysis, the common factor (i.e. depression) is an unobserved (latent) variable that is defined based on observed variables (i.e. the items of the CES-D 8 scale). The common factor is presumed to influence responses to the items (30). CFA thus can be used to test whether this set of items construct an indirect measure of our common factor depression.

Analysis starts with the determination of the best fitting model form of the CES-D 8 scale. In the CES-D 20 literature, the number of factors identified is usually four, namely depressed affect, positive affect, somatic, and interpersonal problems, together loading on the common factor depression (17, 31-36). For the CES-D 8, previous research on the structural form of the scale is not available. However, based on the available items in the 8-item version and the identified structure of the full CES-D, three structural forms can be hypothesised. The first is a one-dimensional model, with all items loading on one common factor depression. An alternative form is a two-dimensional second-order factor model, built up by the factors depressed affect and somatic complaints, each loading on the underlying factor depression (37, 38). Several authors additionally construct a distinct factor of the reversed worded items felt happy and enjoyed life, proposing a three- rather than two-dimensional construct (39). However, we believe that the relationship among the reverse worded items is better accounted for by correlated errors than separate factors. The differential covariance among these items is not based on the influence of a distinct substantially important latent dimension, but rather reflects an artefact of response styles associated with the wording of the items (40, 41).

While CFA allows testing factorial invariance of a construct in a single population group, multigroup CFA (MCFA) measures whether construct validity is invariant across two or more groups. Available tests for multigroup comparison form a nested hierarchy defining several levels of factorial invariance: configural, metric (also called pattern), scalar (also called strong factorial), and residual (also called strict factorial invariance) (42-44). At each level a more restrictive hypothesis is introduced providing increasing evidence of factorial invariance, and allowing specific group comparisons to be made.

Configural invariance - Configural invariance requires that an instrument represents the same number of common factors across groups, and that each common factor is associated with identical item sets across groups. If a specific model form fits well in all groups, then configural invariance is supported. However, configural invariance is not sufficient to defend quantitative group comparisons.

Metric invariance – the hypothesis of metric invariance tests whether the common factors have the same meaning across groups, that is whether the factor loadings are equal across groups. Factor loadings represent the strength of the linear relation between each factor and its associated items (44). When the loading of each item on the underlying factor is equal across groups, the unit of measurement of the underlying factor is identical and the (co)variances of the estimated factors can be compared between groups. Group comparisons are defensible because the meanings of corresponding common factors are deemed invariant across groups and because the MCFA model decomposes total item variation into estimated factor components (i.e. true scores) and residual components (43). Therefore, group differences in common factor variation and covariation are not contaminated by possible group differences in residual variation. If metric invariance is not supported, then two interpretations are possible: On the one hand this might indicate that the meaning of one or more of the common factors, or at least a subset of the items, differs between the groups. On the other hand it might point to an extreme response style by one of the groups.

Scalar invariance – Once the hypotheses of configural and metric invariance are supported, a test of scalar invariance is in order. Such a test addresses the question whether there is differential additive response bias (29, 45, 46). Such bias is caused by forces – such as cultural norms – which are unrelated to the common factors, but systematically cause higher- or lower-valued item response in one population group compared to another. Within the MCFA model, systematic additive influences on responses are reflected in the item intercepts. Since this response style is additive, it affects observed means but not response variation. According to Gregorich (43) evidence that corresponding factor loadings and item intercepts are invariant across groups suggests that 1) group differences in estimated factor means will be unbiased and 2) group differences in observed means will be directly related to group differences in factor means and will not be contaminated by differential additive response bias.

Residual invariance - For most researchers comparison of group means is of main interest. Therefore the highest level of factorial invariance, namely residual invariance, is of limited practical value. Residual invariance allows comparisons of observed variance or covariance across groups. The comparison is defensible because it entirely reflects common factor variation without being contaminated by differences in residuals. It is tested by constraining the residuals associated with each item to be equal across groups, in addition to the loadings and the intercepts of the model.

Measurement invariance of any of the above-mentioned hypotheses is said to be ‘full’ when all parameters are invariant across groups. However, in practical applications, full measurement invariance frequently does not hold. The researcher should then ascertain whether there is at least partial measurement invariance (29). It assumes that the construct is configurally invariant across groups, and that a substantial number of parameters is also invariant in the additional hypotheses. Finding partial invariance suggests that the substantive group comparisons associated with the corresponding ‘full’ invariance hypotheses are defensible since only the subset of items meeting the metric or scalar invariance criteria are used to estimate associated group differences (47).

In summary, measurement or factorial variance can be established at different levels, contaminating estimates of latent constructs in several ways. MCFA allows to compare the means and variance of latent constructs by correcting for possible bias due to variation across groups in the number of common factors and the item/factor clusters (configural invariance), factor loadings (metric invariance), item intercepts (scalar invariance) and residual variances (residual invariance).

Methods

Data: The European Social Survey (ESS) 2006/2007

The European Social Survey (ESS3) (http://www.europeansocialsurvey.org) (16) is a biennial survey covering more than 25 European countries in 2006 and 2007. The ESS is designed to chart and explain the interaction between Europe's changing institutions and the attitudes, beliefs and behaviour patterns of its diverse populations. In each participating country, the ESS-sample is designed following a strict randomised probability procedure and data are gathered with face-to-face interviews. The use of proxies was not allowed. ESS information is representative for all individuals in the general population aged 15 and older, living in a private household, irrespective of their language, citizenship and nationality. In our analyses, we restrict ourselves to the Belgian sample, consisting of 838 male and 956 female respondents. Response rate for the Belgian sample was 61.01%.

The CES-D 8

The Center of Epidemiological Studies Depression Scale or CES-D (17) is a key instrument in the measurement of depression in American research (39), but less often implemented within the European context. In Belgium, the CES-D has not yet been used on a large scale, except in the Epidemiology Research on Dementia in Antwerp (ERDA)-survey, restricted to elderly people (48). Initially, the CES-D was built by 20 self-report items in order to identify populations at risk of developing depressive disorders; in itself however, it should not be used as a clinical diagnostic tool (17). The 20 items primarily measure affective and somatic dimensions of depression, especially reflected in complaints such as depressed mood, feelings of guilt and worthlessness, helplessness and hopelessness, psychomotor retardation, loss of appetite, and sleep disturbance. Respondents are asked to indicate how often in the week previous to the survey they felt or behaved in a certain way ranging from 'none or almost none of the time' to 'all or almost all of the time'. The response values are 4-point Likert scales, with range 0 to 3. Scale scores for the CES-D are assessed using non-weighted summated rating and range from 0 to 60 for the CES-D 20 and from 0 to 24 for the CES-D 8, with higher scores indicating a higher frequency of depressive complaints. International literature shows the CES-D 20 to have good psychometric properties (17, 36, 49). Shorter versions of the CES-D 20 have been used extensively before, but research based on the 8-item version is rather scarce. Based on our Belgian sample, we can confirm the reliability of the CES-D 8 for measurement of depression within a general population context. Reliability was indicated by a response rate of 99.9% in both men and women, and a Cronbach alpha of 0.82 in men and 0.84 in women. The items building up the CES-D 8 are reported in Table 1.

TABLE 1: Items of the 8-item version of the Center of Epidemiological Studies-Depression Scale (51)

How much of the time during the past week…

Answers range from 0 (none or almost none of the time) to 3 (all or almost all of the time)

… did you feel depressed?
… did you feel everything you did was an effort?
… was your sleep restless?
… were you happy?
… did you feel lonely?
… did you enjoy life?
… did you feel sad?
… were you unable to get going?

Statistical procedure

In order to compare depression scores across gender, the CES-D 8 scale requires factorial invariance in both model form and model parameters (44). We estimated the best fitting model using CFA. This model is fitted to male and female data via multigroup analysis using Maximum Likelihood estimations. Analysis is conducted using the AMOS 16.0 programme. The analysis follows two phases: First, measurement invariance is hierarchically tested at each of the levels: dimensional, configural, metric, intercept and residual invariance. Second, we estimate the factor means and variances of the depression construct for both men and women separately. We then compare the estimated mean differences of our factor model with the observed mean differences of men and women.

In our invariance tests, four specific model fit indicators are used. Commonly used in multi-group analyses, is the Chi-square test, testing the magnitude of the discrepancy between the sample and fitted covariance matrices (50). When Chi-square is significant, the model is rejected. However, the Chi-square test may easily lead to a type I error (and thus to an incorrect rejection of the model) in case of non-normality of data, large sample sizes and complex models (see Bollen (51) for a detailed explanation of the influence of sample size on measures of model fit). Since all three conditions are inherent in our study, we report the Chi-square test but add three model fit indices that showed a more robust performance in a simulation study by Hu and Bentler (52): the Tucker-Lewis index (TLI) (53), the Comparative Fit Index (CFI) (54) and the Root Mean Squared Error of Approximation (RMSEA) (55). The first two indices range from 0 (poor fit) to 1 (perfect fit). A value of 0.90 or higher provides evidence for a good fit, a value of 0.95 or above for an excellent fit (52). The RMSEA indicates a reasonable fit in case its score is 0.08 or less and a good fit in case the score is 0.05 or less (56). We additionally evaluate the difference in fit between the more restricted model and the less restricted model by examining the changes in the CFI index. Cheung and Rensvold (57) claim that changes in CFI of -0.01 or less indicate that the invariance hypothesis should not be rejected.

Goodness of fit is further verified by the absence of large modification indices (MI) and expected parameter changes (EPC), which both indicate specific points of ill fit in the model. The MI of a parameter is a conservative estimate of the decrease in chi-square that would occur if the parameter was relaxed (58). The EPC values provide an estimate of how much the parameter is expected to change in a positive or negative direction if it were freely estimated (41). A specific parameter is relaxed only if its MI is highly significant both in magnitude and in comparison with the majority of other MIs and if its EPC is substantial.

When testing all levels of factorial invariance mentioned above, and assessing depression prevalence in men and women in Belgium in the second phase of our analyses, parameter estimates are weighted using the ESS 3-design weight for Belgium in order to correct for differential selection probability.

Results

Tests of factorial invariance hypotheses

The first panel of Table 2 gives an overview of the goodness-of-fit indices of the proposed factor models. The best fitting model of the CES-D 8 instrument is assessed with the pooled dataset by respectively fitting a one- and a two-dimensional model. The analysis is repeated by additionally controlling for measurement effects of the reversed worded items ‘felt happy’ and ‘enjoyed life’. All models are identified by constraining the factor loading of the item ‘felt depressed’ to 1 and its intercept to 0. Our results indicate that all models have a significant chi-square, but the three other indices show only a good fit for the models with correlated errors terms: TLI and CFI above 0.90, RMSEA below 0.08. However, looking closer at the estimates of the two-factor model (results not shown here) indicates that this model includes a negative error variance, making its solution unacceptable. Based on these results we use model 1c – with all items loading on one dimension and with correlated errors between the reversed worded items – as our baseline model for the upcoming MCFA.

TABLE 2: Model fit summary: Chi-square, CFI, TLI and RMSEA. European Social Survey, Belgian sample 2006-2007 (51)

Model	χ^2	df	CFI	TLI	RMSEA
Best fitting model					
1a. One-dimensional	480,795	20	0.898	0.857	0.113
1b. Two-dimensional	354,978	19	0.925	0.890	0.099
1c. One-dimensional – corr. errors	219,309	19	0.955	0.934	0.077
1d. Two-dimensional – corr. errors	101,437	18	0.981	0.971	0.051
MCFA Equivalence tests					
2. Configural	251,921	38	0.951	0.928	0.056
3. Metric	279,331	45	0.946	0.933	0.054
4a. Scalar	315,334	52	0.940	0.935	0.053
4b. Partial scalar	300,134	51	0.943	0.937	0.052
5. Partial residual	331,647	59	0.938	0.941	0.051

The second panel of Table 2 shows the fit statistics of the MCFA across Belgian men and women simultaneously. The results of model 2 indicate that our baseline model fits well in both the male and female sample, providing evidence for configural invariance. The assumption that factor loadings are identical in both groups is also granted based on the findings in model 3. Although the metric invariance test shows a significant chi-square ($\Delta\chi(7) = 27{,}410$, $p < 0.001$), both the CFI and TLI suggest a good fit as indicated by a score above 0.90, with a decrease of less than 0.01. In addition, the RSMEA is smaller than 0.08, adding evidence for metric invariance of the scale across gender. Absence of specific points of ill fit in the model is provided by low MIs and EPCs. Our results therefore indicate that comparing the latent (co)variances of the CES-D 8 across Belgian men and women is valid.

The fourth model in Table 2 tests scalar invariance by additionally imposing equality constraints on corresponding item intercepts. All model fit indices, except for the significant chi-square ($\Delta\chi(7) = 36{,}003$, $p < 0.001$), suggest that the model shows scalar invariance across gender in the general population in Belgium. However, examination of the MIs and EPCs reveals that the intercept of the item ‘felt sad’ is significantly higher in Belgian women compared to men. In order to test partial scalar invariance, we therefore relaxed this intercept. Even though chi-square is still significant ($\Delta\chi(6) = 20{,}803$, $p < 0.001$), CFI and TLI were above 0.90 and RMSEA below 0.08, supporting the hypothesis of partial scalar invariance. These findings suggest that comparisons across gender of factor and observed means of the CES-D 8 are defensible.

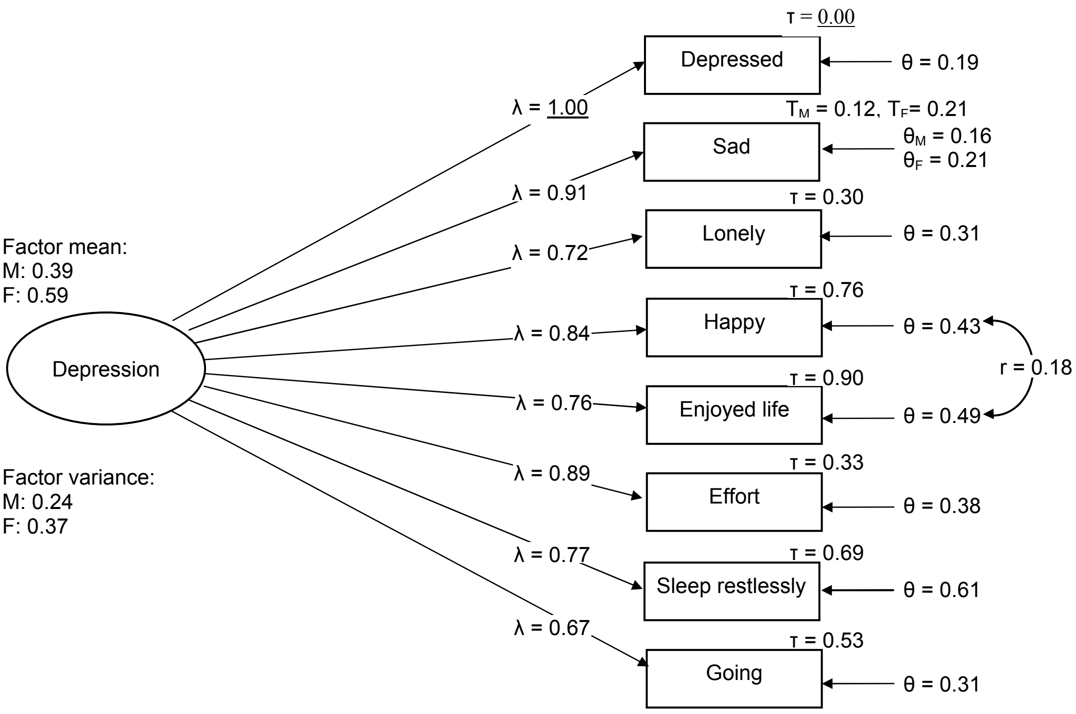
In order to legitimately compare the observed (co)variances of depression in Belgian men and women, the highest level of factorial invariance needs to be verified. This is tested by additionally constraining all corresponding item residual variances, except for the scalar invariant item ‘felt sad’. Model 5 shows that the hypothesis of partial residual invariance is empirically supported. Chi-square is again significant ($\Delta\chi(8) = 30{,}387$, $p < 0.001$), but both CFI and TLI are above 0.90 and RMSEA below 0.08. No significant MIs and EPCs could be found, confirming factorial invariance at all levels.

In sum, the findings for all models in Table 2 indicate that the one-dimensional CES-D 8 scale with correlated errors between the positively worded items ‘felt happy’ and ‘enjoyed life’ can be used to compare factor means and (co)variances and observed means and (co)variances between men and women in the general population in Belgium. The absence of a decrease in CFI of more than 0.01 with each more restrictive model additionally provides evidence for factorial invariance at all levels. The final model is depicted in Figure 1. The parameter estimate subscripts identify sample membership (i.e. M for male or F for female). Model-identifying constraints are underlined. As the figure shows, all observed items load on the latent construct depression with factor loadings of at least $\lambda=0.67$. Intercepts are equal across gender for all items, except for ‘felt sad’, with women showing a systematically higher-valued item response pattern indicated by a higher unrestricted parameter estimate of $T_F= 0.21$ compared to $T_M = 0.12$, for men. Corresponding unrestricted residuals were $\theta_M = 0.21$

and θ_M = 0.16 respectively. The correlation between the residuals of the positively worded items is shown to be r=0.18.

FIGURE 1: Partial residual invariance model of CES-D 8 model for male and female data

European Social Survey, Belgian sample 2006-2007 (51)



λ = factor loading estimate, τ = factor/item intercept estimate, θ = residual variance estimate, r = residual covariance estimate
M = male, F = female

Comparison of observed means and variances

The factorial invariance tests reported in Table 2 showed empirical evidence for the hypotheses of partial residual invariance, indicating that comparisons of observed means and variances of the CES-D 8 in men and women is warranted. In Table 3 the observed means and variances on all composite items and on the total CES-D 8 scale score are reported. Because the partial residual invariance model holds, we expect that observed group differences in means and variances will be similar to corresponding group differences in factor means and variances.

Female respondents score significantly higher on all observed items of the CES-D 8. The high MI and EPC of the item ‘felt sad’ predicted a significant difference between men and women. This is confirmed by the observed means with largest gender difference for this item.

TABLE 3: Comparison of observed means and variances with estimated factor means and variance European Social Survey, Belgian sample 2006-2007 (51)

	MEANS			VARIANCES		
	Male	Female	Δ; p	Male	Female	Ratio; p
Observed score						
Depressed	0.36	0.61	0.25; 0.000	0.38	0.59	0.64; 0.000
Sad	0.37	0.65	0.28; 0.000	0.35	0.52	0.67; 0.000
Lonely	0.32	0.43	0.11; 0.001	0.41	0.52	0.79; 0.000
Happy	0.95	1.06	0.11; 0.004	0.63	0.68	0.93; 0.185
Enjoyed life	0.95	1.10	0.15; 0.000	0.63	0.70	0.90; 0.050
Everything an effort	0.59	0.73	0.16; 0.000	0.54	0.69	0.78; 0.005
Restless sleep	0.73	0.93	0.20; 0.000	0.70	0.87	0.80; 0.132
Could not get going	0.49	0.56	0.07; 0.032	0.48	0.53	0.91; 0.031
Mean total score	0.59	0.76	0.17; 0.000	0.22	0.29	0.76; 0.000
Estimated score	0.39	0.59	0.20; 0.000	0.24	0.37	0.65; 0.000

The gender difference is smallest for the item ‘could not get going’. In both groups the positively worded items ‘felt happy’ and ‘enjoyed life’ have the highest scores, while ‘feeling lonely’ occurs least in both men and women. The overall mean of the CES-D 8 also differs significantly in the male and female sample, with a difference of 0.17. Our observed results thus point to a higher prevalence of depressive symptoms in the female sample. Comparisons of the item variances suggest significant group differences for all items except the items ‘felt happy’, ‘enjoyed life’ and ‘restless sleep’ with larger variances of the item scores and overall score for women than men. So even though women on average score higher than men, their scores are more spread out than those of male respondents.

Based on the partial residual model shown in Figure 1 we estimated a difference in factor means between the two groups of 0.20, which is slightly larger than the 0.17-difference of the observed means (the difference amounts to 11% of the total sample standard deviation of 0.52). Similarly we note a reasonably small difference in the ratio of the estimated versus observed variance (0.76 in observed variance versus 0.65 in estimated variance). As expected our estimated scores correspond closely to our observed scores.

Discussion

Simultaneous analysis of multiple groups places higher demands on the measurement scale than single-group research. It requires that instruments measure constructs with the same meaning across groups and allow defensible quantitative group comparisons. In this study, we used a scale that measures depression by assessing the frequency and occurrence of certain depressive symptoms. Depression is thus considered to be a latent construct whose properties are inferred from observing the set of variables that serve as manifest indicators. If the mean depression score of men differs from that of women, what can be concluded? It may be the case that these two groups actually differ in their level of depression, it may also be the case that extraneous influences are giving rise to the observed differences (30).

Therefore factorial invariance needs to be tested in quantitative comparative research. Unfortunately, factorial invariance has been tested relatively infrequently in past research. When it is tested, researchers predominantly focused on invariance of the factor construct (59). The number of studies that determine whether comparisons of group means are defensible has increased in recent years, but its application is not yet widely used and scattered across different domains. Lack of requisite technical skills or lack of awareness might explain the scarceness of these types of invariance studies (43).

In the present study we established factorial invariance at all levels: dimensional, configural, metric, scalar and residual invariance. Next we estimated depression mean scores and variances across gender, eliminating a measurement artefact. Our results indicate that the CES-D 8 scale can be used to compare mean differences in depression in men and women, showing good reliability and validity. Our study based on the ESS 3 data of the general population in Belgium confirms the consistent epidemiological finding that women report more complaints of depression than men. Moreover, the analyses show that compared to men, women score higher on all the items of the CES-D 8. Although the difference between the observed and estimated gender difference in depression is small, our results suggest that the true gender difference in depression is somewhat larger than the observed answers of Belgian women and men on the 8-item short version of the CES-D

Some limitations of our study are worth noting when interpreting the results. When testing factorial invariance in large community samples such as the ESS 3, the researcher should also bear in mind that the variables of interest are often non-normally distributed, specifically when working with ordinal Likert scales (60). However, the maximum likelihood estimation method assumes that data have a normal distribution. In our analyses, we tested the robustness of our findings by additionally estimating a Bollen-Stine significance level via bootstrapping, a procedure compensating for the normality assumption (61). Results (not shown) did not indicate a different significance level than the one reported for the Chi-square tests. An additional robustness test was based on a logarithmic transformation of the CES-D 8 data, decreasing the non-normality of the item and scale score distributions. This procedure results in better fit-indices (not shown), but it simultaneously increases the complexity of a substantive interpretation of the parameter estimates. Important to note is that the hypotheses of factorial invariance were supported by all estimation methods even after controlling for non-normality.

Finally, the current findings do not automatically imply psychometric equivalence across gender outside the Belgian context, or across social groups distinguished by other criteria such as language, ethnicity, social class or age. All these social groups may have group-specific attributes that lead to measurement inequivalence of (self-report) scales. We therefore strongly suggest to test factorial invariance before comparing specific group scores. In the Belgian context for example, it might be relevant to examine a possible language bias. In our opinion, the Dutch and French translations are insufficiently equivalent to the original English scale. Especially the translation of the item ‘could not get going’ deserves specific attention[2]. The actual experience and expression of depression may vary sufficiently according to other demographic and social or cultural factors to effectively undermine attempts to compare rates of depressive symptoms across all groups. Further research is needed to determine the extent to which these factors influence responses to self-report instruments.

Conclusion

The CES-D 8 can be considered a reliable and valid measurement instrument for depression within the general population context in Belgium. A three-dimensional depression model, built up by the factors 'depressed affect', 'positive affect' and 'somatic' fits the data best. Measurement equivalence tests show that the scale allows defensible cross-gender comparisons leading to prevalence estimates that are not contaminated by group-specific elements unrelated to depression. Consistent with international literature, we found higher levels of depression in women compared to men, but our analyses suggest that the true gender difference in depression is somewhat larger than the one observed in the ESS 3.

Acknowledgements

The research reported in this article was supported by a special research grant from Ghent University (Bijzonder Onderzoeksfonds 2007).

References

1. WHO. Women's Mental Health. 2000.
2. Weissman MM, Klerman GL. Sex-Differences and Epidemiology of Depression. Arch Gen Psychiatry 1977; 34(1): 98-111.
3. Weissman MM, Leaf PJ, Holzer CE, Myers JK, Tischler GL. The Epidemiology of Depression - An Update on Sex-Differences in Rates. J Affect Disord 1984; 7(3-4): 179-88.
4. Kessler RC, Mcgonagle KA, Swartz M, Blazer DG, Nelson CB. Sex and Depression in the National Comorbidity Survey .1. Lifetime Prevalence, Chronicity and Recurrence. J Affect Disord 1993; 29(2-3): 85-96.
5. Piccinelli M, Wilkinson G. Gender differences in depression - Critical review. Br J Psychiatry 2000; 177: 486-92.
6. Ansseau M, Fischler B, Dierick M, Mignon A, Leyman S. Prevalence and impact of generalized anxiety disorder and major depression in primary care in Belgium and Luxemburg: the GADIS study. Eur Psychiatry 2005; 20(3): 229-35.
7. Lepine JP, Gastpar M, Mendlewicz J, Tylee A. Depression in the community: The first pan-European study DEPRES (Depression Research in European Society). Int Clin Psychopharmacol 1997; 12(1): 19-29.
8. Tylee A, Gastpar M, Lepine JP, Mendlewicz J. Identification of depressed patient types in the community and their treatment needs: findings from the DEPRES II (Depression Research in European Society II) Survey. Int Clin Psychopharmacol 1999; 14(3): 153-65.

[2] Translated in French as: 'd'être incapable de rien faire', in Dutch as 'u niet op gang kon komen / niet op gang kon komen in de zin van "u voelde zich (s)loom en niet gemotiveerd"'.

9. Alonso J, Angermeyer MC, Bernert S, Bruffaerts R, Brugha IS, Bryson H, et al. Prevalence of mental disorders in Europe: results from the European Study of the Epidemiology of Mental Disorders (ESE-MeD) project. Acta Psychiatr Scand 2004; 109: 21-7.
10. Bruffaerts R, Bonnewyn A, Van Oyen H, Demarest S, Demyttenaere K. Prevalentie van mentale stoornissen in de Belgische bevolking. Resultaten van de European Study on Epidemiology of Mental Disorders (ESEMeD). Tijdschr Geneeskd 2004; 60: 75-85.
11. Bracke P. The three-year persistence of depressive symptoms in men and women. Soc Sci Med 2000; 51(1): 51-64.
12. Bracke P. Sex differences in the course of depression: evidence from a longitudinal study of a representative sample of the Belgian population. Soc Psychiatry Psychiatr Epidemiol 1998; 33(9): 420-9.
13. Bracke P. Geslachtsverschillen in depressief gedrag in een representatieve steekproef van de Vlaamse bevolking: de validiteit van een zelfrapportageschaal. Archives of Public Health 1996; 54: 275-300.
14. Levecque K. Generalized anxiety and depression in the general population: Risk factors according to the Belgian Health Interview Survey 2001. Depress Anxiety 2006; 23(8): 509-11.
15. Levecque K, Lodewyckx I, Vranken J. Depression and generalised anxiety in the general population in Belgium: A comparison between native and immigrant groups. J Affect Disord 2007; 97(1-3): 229-39.
16. Jowell R. European Social Survey 2006/2007. Round 3: Technical Report. London: Centre for Comparative Social Surveys, City University, 2007.
17. Radloff LS. The CES-D Scale: A self-report depression scale for research in the general population. Applied Psychological Measurement 1977; 1: 385-401.
18. Callahan CM, Wolinsky FD. The Effect of Gender and Race on the Measurement Properties of the Ces-D in Older Adults. Med Care 1994; 32(4): 341-56.
19. Cole SR, Kawachi I, Maller SJ, Berkman LF. Test of item-response bias in the CES-D scale: Experience from the New Haven EPESE study. J Clin Epidemiol 2000; 53: 285-9.
20. Stommel M, Given BA, Given CW, Kalaian HA, Schulz R, Mccorkle R. Gender Bias in the Measurement Properties of the Center-For-Epidemiologic-Studies-Depression-Scale (Ces-D). Psychiatry Res 1993; 49(3): 239-50.
21. Ross CE, Mirowsky J. Components of Depressed Mood in Married Men and Women - the Center for Epidemiologic Studies Depression Scale. Am J Epidemiol 1984; 119(6): 997-1004.
22. Ross CE, Mirowsky JJ. Socially-desirable response and acquiescence in a cross-cultural survey of mental health. J Health SocBehav 1984; 25: 189-97.
23. Berkman LF, Berkman CS, Kasl S, Freeman DH, Leo L, Ortfeld AM, et al. Depressive symptoms in remation to physical health and functioning in the elderly. Am J Epidemiol 1986; 124: 372-88.
24. Clark VA, Aneshensel CS, Frerichs RR, Morgan TM. Analysis of Effects of Sex and Age in Response to Items on the Ces-D Scale. Psychiatry Res 1981; 5(2): 171-81.
25. Byrne BM, Baron P, Campbell TL. Measuring Adolescent Depression: Factorial Validity and Invariance of the Beck Depression Inventory Across Gender. J Res Adolesc 1993; 3: 127-43.
26. Mirowsky JJ, Ross CE. Sex differences in distress - Real or artifact. Am Sociol Rev 1995; 60(449): 468.
27. Nunally J. Psychometric theory. New York: McGraw-Hill, 1978.
28. Meredith W. Measurement invariance, factor-analysis and factorial invariance. Psychometrika 1993; 58: 525-43.
29. Steenkamp JBEM, Baumgartner H. Assessing measurement invariance in cross-national consumer research. J Consum Res 1998; 25: 78-90.
30. Meredith W. An essay on measurement and factorial invariance. Med Care 2006; 44: S69-77.
31. Golding JM, Aneshensel CS. Factor structure of the Center of Epidemiologic Studies Depression Scale among Mexican Americans and Non-Hispanic Whites. J Consult Clin Psychol 1989; 1: 163-8.

32. Hertzog C, Van Alstine J, Usala PD, Hultsch DF, Dixon R. Measurement properties of the Center for Epidemiological Studies Depression Scale (CES-D) in older populations. Psychol Assess 1990; 2(1): 64-72.
33. Joseph S, Lewis CA. Factor-Analysis of the Center for Epidemiologic Studies-Depression Scale. Psychol Rep 1995; 76(1): 40-2.
34. Roberts RE, Vernon SW, Rhodes HM. Effects of language and ethnic status on reliability and validity of the Center for Epidemiologic Studies-Depression Scale with psychiatric patients. J Nerv Ment Dis 1989; 177: 581-92.
35. Roberts RE, Rhoades HM, Vernon SW. Using the Ces-D Scale to Screen for Depression and Anxiety - Effects of Language and Ethnic Status. Psychiatry Res 1990; 31(1): 69-83.
36. Shafer AB. Meta-analysis of the factor structures of four depression questionnaires: Beck, CES-D, Hamilton, and Zung. J Clin Psychol 2006; 62(1): 123-46.
37. Riddle AS, Hess U. Static versus dynamic structural models of depression: The case of the CES-D. 2008.
38. Steffick D. Documentations of affective functioning measures in the Health and Retirement Study. HRS/AHEAD Documentation Report DR-005. Ann Arbor, MI: Survey Research Center, University of Michigan, 2000.
39. Perreira KM, Deeb-Sossa N, Harris KM, Bollen K. What are we measuring? An evaluation of the CES-D across race/ethnicity and immigrant generation. Soc Forces 2005; 83(1567): 1601.
40. Marsch HW. Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? J Pers Soc Psychol 1996; 70: 810-9.
41. Brown TA. Confirmatory factor analysis in applied research. New York: The Guildford Press; 2006.
42. Byrne BM. Multigroup Comparisons and the Assumption of Equivalent Construct Validity Across Groups: Methodological and Substantive Issues. Multivariate Behav Res 1989; 24: 503-23.
43. Gregorich SE. Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. Med Care 2006; 44: S78-94.
44. Bollen KA. Structural equations with latent variables. New York: Wiley, 1989.
45. Rorer LG. The great response-style myth. Psychol Bull 1965; 63: 123-56.
46. Cheung GW, Rensvold RB. Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. J Cross Cult Psychol 2000; 31: 187-212.
47. Byrne BM, Shavelson RJ, Muthén B. Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. Psychol Bull 1989; 105: 456-66.
48. Van Oyen H. Epidemiology research on dementia in Antwerp (ERDA). 1990.
49. Sheehan TJ, Fifield J, Reisine S, Tennen H. The Measurement Structure of the Center for Epidemiologic Studies Depression Scale. J Pers Assess 1995; 64(3): 507-21.
50. Chen FF, Sousa KH, West SG. Testing measurement invariance of second-order factor models. Structural Equation Modeling-A Multidisciplinary Journal 2000; 12: 471-92.
51. Bollen KA. A new incremental fit index for general structural equation models. Sociol Methods Res 1989; 17(303): 316.
52. Hu LT, Bentler PM. Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. Psychol Methods 1998; 3(4): 424-53.
53. Tucker LR, Lewis C. Reliability coefficients for maximum likelihood factor-analysis. Psychometrika 1973; 38: 1-10.
54. Bentler PM. Comparaitive fit indexes in structural models. Psychol Bull 1990; 107: 238-46.
55. Steiger JH. Structural model evaluation and modification – an interval estimation approach. Multivariate Behav Res 1990; 25: 173-80.

56. Browne MW, Cudeck R. Alternative ways of assessing model fit. Sociol Methods Res 1992; 21: 230-58.
57. Cheung GW, Rensvold RB. What constitutes significant differences in evaluating measurement invariance? 1999. [Paper presented at the 1999 conference of the Academy of Management, Chicago]
58. Arbuckle JL. AMOS 16.0 User's Guide. SPSS inc.; 2007.
59. Vandenberg RJ, Lance CE. A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. Organisation Research Methods 2000; 2: 4-69.
60. Lubke GH, Muthen BO. Applying multigroup confirmatory factor models for continuous out-comes to Likert scale data complicates meaningful group comparisons. Structural Equation Modeling 2004; 11: 514-34.
61. Nevitt J, Hancock GR. Relative performance of rescaling and resampling approaches to model Chi-square and parameter standard error estimation in structural equation modeling. San Diego 1997.