Gender differences in depression in 25 European countries after eliminating measurement bias in the CES-D 8

Reference:

Van de Velde Sarah, Bracke Piet, Levecque Katia, Meuleman Bart.- Gender differences in depression in 25 European countries after eliminating measurement bias in the CES-D 8

Social science research: a quarterly journal of social science methodology and quantitative research - ISSN 0049-089X - 39:3(2010), p. 396-404

Full text (Publishers DOI): http://dx.doi.org/doi:10.1016/j.ssresearch.2010.01.002

Gender differences in depression in 25 European countries after eliminating measurement bias in the CES-D 8

Sarah Van de Velde[1], Piet Bracke[1,2], Katia Levecque[1,3], Bart Meuleman[3,4]

Reference: Van de Velde, Sarah, et al. "Gender differences in depression in 25 European countries after eliminating measurement bias in the CES-D 8." *Social Science Research* 39.3 (2010): 396-404.

[1] Department of Sociology, Ghent University,

[2] Corresponding author: Address: Korte Meer 5, 9000 Gent, Belgium. Tel.: +32 9 264 68 03, fax: +32 9 264 69 75. Email address: piet.bracke@ugent.be

[3] Research Foundation (FWO) - Flanders

[4] Centre for Sociological Research, Catholic University Louvain

Abstract

Cross-national comparisons of the prevalence of depression in general populations are hampered by the absence of comparable data. Using information on the frequency and severity of depressive symptoms from the third wave of the European Social Survey (ESS-3), we are able to fill this gap. In the ESS-3, depression is measured with an eight-item version of the Center for Epidemiological Studies-Depression (CES-D 8) scale. Using multigroup confirmatory factor analysis, we assess configural, metric, and scalar measurement invariance of the CES-D 8. Next, best-fitting factor models are used **for latent mean comparisons of** women and men in the 25 participating European countries. The present study is the first to present highly comparable data on the prevalence of depression in women and men in Europe. Results show that, after eliminating measurement bias, the gender difference in depression stays significant and regional clustering can be noted.

Key words: depression, gender, Europe, measurement invariance, CES-D 8

Introduction

Cross-national Gender Differences in Depression: a Measurement Bias?

According to the World Health Organization (WHO), depression is the most common mental health problem in the Western world (WHO 2000). It has a high prevalence in nearly every society; some studies even suggest that depression is on the rise, at least in the Western world (Wauterickx and Bracke 2005; Kessler et al.1993). A recurrent finding in international literature is that there is a 1.5 to 3 times greater prevalence of depression in women compared to men (Piccinelli and Wilkinson 2000; Bebbington 1996). This is true for both inpatient and outpatient studies, as well as for general population studies. The pattern of a higher prevalence of depression in women compared to men is consistent across nations, cultures, and population groups, in studies using different methods and measurement instruments, and for a diversity of incidence and prevalence indicators (Weissman et al.1984; Kessler et al.1993).

Unfortunately, cross-national comparisons of gender differences in depression in the general population have been hampered by the absence of comparable data. Usually, cross-national differences are estimated using meta-analyses of data from a diverse set of studies using divergent depression inventories, different sampling designs, or sampling populations that are not completely comparable. In Europe three cross-national psychiatric epidemiological surveys went beyond those limitations and delivered comparable data: the DEPRES I and II studies, the ESEMeD study, and the ODIN investigation. Nevertheless, these studies too have their drawbacks. The DEPRES II study (Angst et al. 2002) allows for the estimation of gender differences in depression in a cross-national sample of treated individuals only. Moreover, like the DEPRES I study (Lepine et al. 1997), this sample consists of only 6 countries. The ESEMeD study (Alonso et al. 2004) also contains data from only 6 European countries. Finally, the ODIN study (Ayuso-Mateos et al. 2001) contains information on individuals from nine urban centers and rural areas in 5 countries, albeit some of them were identified via primary care databases. Other studies containing information on depression and anxiety-related complaints are (a) the Survey of Health, Ageing and Retirement in Europe (SHARE) and (b) the WHO's Psychological Problems in Primary Care study. SHARE covers 11 European countries, but is limited to couples aged 50 and older (Börsch-Supan et al. 2005). The WHO study sampled clients of

15 primary care centers in 14 countries (Maier et al. 1999). The samples were enriched with depressed cases, so they are not representative for the general population.

In the present study, we make use of the third round of the European Social Survey (ESS-3) (Jowell 2007), organized in 2006 and 2007 and covering data from 25 European countries. In the ESS-3 the frequency and severity of symptoms related to the DSM IV criteria for major depressive disorders are measured using a shorter version of the Center for Epidemiologic Studies-Depression (CES-D) scale (Radloff 1977). Since its introduction the scale has been used to measure depressive symptoms across several populations (elderly, adolescents, women, clinical populations, and ethnic populations). The ESS-3 thus allows us to compare gender differences in depression across multiple European countries.

An analysis of gender and cross-cultural differences in rates of depression, however, presupposes that this concept is measured in an equivalent or invariant way (Moors 2004; vandeVijver 2003). In our study the notion *measurement invariance* refers to “whether or not, under different conditions of observing and studying phenomena, measurement operations yield measures of the same attribute” (Horn and McArdle 1992). If measurement invariance is absent, comparisons across gender or cultural groups become highly problematic. After all, observed between-group differences might be due to measurement artifacts rather than to real differences in the prevalence of depression (Vandenberg and Lance 2000). The CES-D, like most other mental health assessment instruments, was initially developed and tested on samples comprised of mainly European Americans. Most research on the cross-cultural equivalence of the CES-D was done on populations within the United States and, at a later stage, with Hispanic or Asian populations. Unfortunately, validation of the CES-D across European countries is still lacking. A study of the validity of translated versions of the CES-D scale has been made for German populations (Hautzinger 1988), Dutch populations (Bouma et al. 1995), Turkish and Moroccan populations in the Netherlands (Spijker et al. 2004), Portuguese populations (Goncalves and Fagulha 2004), and French populations (Fuhrer and Rouillon 1989). A recent study by Meads, McKenna, and Doward (2006) assessed the comparability of the CES-D scale across UK, German, and US populations and found a bad fit for the European countries.

The results of previous research on the measurement invariance of the scale across gender are also ambivalent, with a number of studies confirming measurement invariance (Berkman et al. 1986; Clark et al. 1981), while other studies point towards a bias (Callahan and Wolinsky 1994;

Stommel et al. 1993). Additionally, several other depression inventories show a gender bias (Byrne, Baron, and Campbell 1993; Mirowsky and Ross 1995;Piccinelli and Wilkinson 2000). In the ESS-3, depression was not assessed using the full-length CES-D scale, but instead respondents were administered an 8-item version, with many of the biased items excluded. It is therefore uncertain whether the shortened version shows measurement invariance properties in cross-cultural and cross-gender research.

Measurement Invariance: a Hierarchy of Hypotheses

Comparative research raises methodological issues that do not present themselves in single-group surveys. In cross-national and gender research, three types of bias can be distinguished: construct bias, item bias, and method bias (vandeVijver and Poortinga 1997). The first type of bias occurs when the construct measured, in this case, depression, is not identical across groups. The definition of depression or the symptoms associated with it might differ slightly across groups, interfering with the equivalence of the measurement. Item bias occurs most often when specific items of the depression scale are translated poorly, when there is low familiarity with the item content in certain cultures, or when there are cultural specifics such as nuisance factors or connotations associated with the item wording **causing extreme or additive response styles** (vandeVijver 2003). Finally, method bias occurs when there are sample incomparabilities, instrument differences, interviewer or respondent effects, or differences in the mode of administration.

In analysis, these forms of bias should be identified as measurement error. However, commonly used approaches such as ordinary least squares assume that variables have been measured without error, “that is, they are perfectly reliable, meaning that all of an observed measure’s variance is true score variance” (Brown 2006). Multigroup confirmatory factor analysis (MCFA) allows for relationships to be estimated after adjusting for certain types of measurement error. In particular, it offers a very strong analytic framework for evaluating the invariance of measurement models across distinct groups (e.g., demographic groups defined by gender and nationalities) and is currently considered the methodology of choice for assessing cross-national measurement invariance (Steenkamp and Baumgartner 1998).

The measurement invariance analysis starts with evaluating the best fitting model form of the CES-D 8 scale. In CES-D 20 literature, the number of factors identified is usually four, namely, *depressed affect*, *positive affect, somatic complaints,* and *interpersonal problems,* that together load on the common factor *depression* (Perreira et al. 2005; Radloff 1977;Golding and Aneshensel 1989;Hertzog C et al. 1990; Joseph and Lewis 1995; Shafer 2006). 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 hypothesized. 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* (Riddle and Hess 2008; Steffick 2000). Several authors additionally construct a distinct factor of the reverse-worded items *were happy* and *enjoyed life*, proposing a three- rather than two-dimensional construct (Perreira et al. 2005). 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 artifact of response styles associated with the wording of the items (Brown 2006; Marsch 1996).

After identifying the best fitting structural form for the measurement model, we use MCFA to evaluate the cross-national and cross-gender measurement invariance of the construct. The available tests for MCFA form a nested hierarchy defining several levels of measurement invariance: configural, metric, and scalar invariance (Bollen 1989; Byrne 1989; Meredith 1993). At each level a more restrictive hypothesis is introduced, providing increasing evidence of measurement 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 groups comparisons.

Metric invariance. The hypothesis of metric invariance tests whether the corresponding factor loadings are equal across groups. 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. **Metric invariance can be disturbed by extreme response styles (ERS) due to culturally based response norms, which cause bias to a**

subset of the factor loadings in one or more groups (Gregorich 2006). Respondents in high-ERS countries might favor decisiveness or humility, whereas in low-ERS countries respondents might desire to appear modest and non-judgmental (Cheung and Rensvold 2000). Some cultures may have very strong opinions about certain topics or feelings, while others may have no opinion (Riordan and Vandenberg 1994). In both cases the first group will favor extreme categories in self-report scales, while the second group will tend to cluster around middle categories, In the case of a self-report scale such as the CES-D 8, certain groups thus might restrain or overrate their reported of level of depression. This results in cross-cultural differences in response style, unrelated to the construct of interest.

Scalar invariance. Scalar invariance is tested by restricting the corresponding item intercepts so that they are equal across groups. **This level of measurement invariance addresses the question of whether there is differential additive response style (ARS) bias (Cheung and Rensvold 2000;Rorer 1965) which results in systematically higher- or lower-valued item responses in one population group compared to another. Cross-cultural ARS differences have been demonstrated in previous studies (Baumgartner and Steenkamp 2001; Cheung and Rensvold 2000). For example, a study by Riordan and Vandenberg (1994) indicated that the middle answer category of a Likert-scale had a different meaning in different cultures. Within the CFA model, ARS is reflected in the item intercepts.** When this level of measurement invariance is met, the group comparisons of latent and observed means are valid.

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* (Steenkamp and Baumgartner 1998), which assumes that the construct is configurally invariant across groups, and that a substantial number of parameters are 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 (Byrne, Shavelson, and Muthén 1989).

In the current study we aim to determine whether the CES-D 8 scale is psychometrically equivalent across gender and countries involved in the ESS-3 by testing its measurement invariance.

The different hypotheses of measurement invariance have been tested relatively infrequently in the past (Gregorich 2006). When they have been tested, investigators have predominantly focused on invariance of construct validity (Vandenberg and Lance 2000) or on the identification of biased items from the full 20-item scale and the reduction of its length (Cole et al. 2000; Perreira et al. 2005; Vega and Rumbaut 1991). The current study makes use of an abbreviated CES-D scale and tests whether its equivalence can effectively be determined across gender and countries in the ESS-3. The aim of our study is therefore threefold. First, we determine the best fitting model for our data. This baseline model is then used to test the hierarchal hypotheses of measurement invariance, allowing us to control for gender and cultural bias in the measurement of depression. Finally, we use our model to estimate latent means of the CES-D 8 and compare these to the observed means. To the best of our knowledge, the present study is the first to present highly comparable data on gender differences in the prevalence of depression in Europe.

Materials and Methods

Sample

Our analyses use data from the third round of the European Social Survey (ESS-3) (Jowell 2007), which covered 25 European countries in 2006 and 2007. For each participating country, respondents were selected by means of strict probability samples of the resident population aged 15 years and older living in private households (irrespective of nationality or language). The use of proxies was not allowed. Data was gathered via face-to-face interviews. Response rates range from 45.97% in France to 73.19% in Slovakia. After deleting cases lacking minimum information on gender or depression, our unweighted sample consists of 46,669 respondents, of which 45% are male.

The CES-D 8

The Center of Epidemiological Studies-Depression (CES-D) scale (Radloff 1977) is a key instrument in the measurement of depression in American research, but is implemented less often within the European context. **Initially, the CES-D was built using 20 self-report items in order to identify populations at risk of developing depressive disorders; it should not, however, be used as a clinical diagnostic tool by itself (Radloff 1977). The ESS-3 includes 8 items from the original scale. Respondents are asked to indicate how often in the week previous to the survey**

they 1) felt depressed, 2) felt lonely, 3) felt sad, 4) were happy, 5) enjoyed life, 6) felt everything they did was an effort, 7) had restless sleep, and 8) could not get going. Response options range from *none or almost none of the time* (score 1) to *all or almost or all of the time* (score 4). Scale scores are assessed using a non-weighted summated rating and range from *8* to *32*, with higher scores indicating a higher intensity of depressive complaints.

Based on the data of the ESS-3, we can confirm the reliability of the CES-D 8 for measurement of depression within a general population context. Total response rates ranged between 95% in men and 94% in women, with the lowest in Ukraine (77.1%) and the highest in Norway (99.8%). Respondent mean substitution was applied to respondents answering at least 5 items of the scale. Respondents who answered fewer than 5 items of the CES-D 8 scale (330 cases) or who did not report their gender (100 cases) were excluded from our analysis. The Cronbach alpha of the CES-D 8 scale was 0.812 in male data and 0.847 in female data with the lowest score in Denmark (0.728) and highest in Hungary (0.881). Consistent with international literature, our findings show higher levels of depression in women compared to men, with a mean score of 14.8 in women and 13.7 in men ($F_{(1, 15147.037)}$: 834.724, $p < 0.001$).

Statistical Procedure

Measurement invariance is examined via MCFA using maximum likelihood estimations. Analysis is conducted using the AMOS 16.0 program. We evaluate the acceptability of our model on the basis of overall goodness of fit in additional to specific points of ill fit. The standard way to compare the overall fit of the different models is the chi-square test. However, this 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. Since the first two conditions are inherent to our study, we also report three model fit indices that have shown a more robust performance (Hu and Bentler 1998): the Tucker-Lewis index (TLI) (Tucker and Lewis 1973), the Comparative Fit Index (CFI) (Bentler 1990), and the Root Mean Squared Error of Approximation (RMSEA) (Steiger 1990). 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, and a value of 0.95 or above for an excellent fit (Hu and Bentler 1998). The RMSEA indicates a reasonable fit when its score is 0.08 or less and a good fit when the score is 0.05 or less (Browne and Cudeck 1992).

Goodness of fit is further verified by the absence of large modification indices (MIs) 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 (Arbuckle 2007). The EPC values provide an estimate of how much the parameter is expected to change in a positive or negative direction if freely estimated (Brown 2006). 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.

Results

Tests of Measurement Invariance Hypotheses

Table 1 gives an overview of the goodness-of-fit indices of the different levels of measurement invariance. In a first step the best fitting model of the CES-D 8 instrument is assessed with the pooled dataset by respectively fitting a one- and two-dimensional model (Model 1a-1b) to our data. The analysis is repeated by additionally controlling for measurement effects of the reverse-worded items *were happy* and *enjoyed life* (Model 1c-1d)[5]. All models are identified by constraining the factor loading of the item *felt* **depressed** to 1 and its intercept to 0. As shown in the first panel of Table 1, all models have a significant chi-square, but the three other indices show only a good fit for the models with correlated-error terms—TLI and CFI above 0.90, RMSEA below 0.08. **However, the two dimensions of model 1d correlate strongly (0,91), making their discriminant validity problematic (Cohen et al.2003).** Based on these results we use Model 1c—with all items loading on one dimension and with correlated errors between the reverse-worded items—as our baseline model for the upcoming MCFA.

The second panel of Table 1 shows the fit statistics of the MCFA, with 50 groups defined by gender and country simultaneously. As Model 2 in Table 1 shows, imposing equality constraints on the underlying factor and item sets provides evidence for configural invariance of the CES-D 8 across

[5] **Reverse-worded items might cause certain respondents to be more inclined to choose for one side of the answer scale, irrespective of the contents of the item. This response style strengthens correlations between items that are formulated in the same direction, and weakens relations between oppositely worded items (Billiet and McClendon2000). Regardless of this method effect, these items are retained in the scale, since excluding them would reduce the substantive value of the scale, and previous research did not indicate them to be problematic (Perreira et al.2005). An error correlation is however allowed between the two reverse-worded to account for this artifact.**

gender and countries. The assumption that factor loadings are identical (metric invariance) in both the male and female data from the different countries is also supported based on the overall goodness-of-fit indices. Although there was a significant decrease in chi-square between the model of configural and metric invariance ($\Delta\chi(343) = 1{,}868.73$, $p < 0.001$), the alternative indices determine a good fit, with CFI and TLI above 0.90 and RMSEA below 0.05. In addition, examination of the MIs and EPCs reveals no specific points of ill fit. Our results therefore indicate that comparing the latent (co)variances of the CES-D 8 across gender and countries is valid.

Insert Table 1 here

The fourth model in Table 1 tests scalar invariance by additionally imposing equality constraints on the corresponding item intercepts. We reject this model; the increase in chi-square is significant ($\Delta\chi(392) = 14{,}564.723$, $p < 0.001$), and all fit indices except RMSEA are below the acceptable level. Based on the MIs and EPCs, we relaxed our model restrictions in order to meet partial scalar invariance. Of all 400 constrained intercepts in our model, **83** needed to be freed for the model to show an acceptable fit and for specific points of ill fit to be eliminated. The relaxed intercepts of the German **males** and **French females** contribute most to the increase in fit. On the other hand, the Belgian, Estonian, Irish, **and Slovenian** populations show very stable conditions with no significant MIs reported. The EPCs indicate that of the 85 intercepts freed, a little less than half were expected to be lower compared to the other groups. The Northern European countries also have lower intercepts than expected, and the Eastern European countries have higher intercepts than expected. This suggests that certain regional cultural norms cause systematic lower-valued item responses in the Northern European countries while in the Eastern European countries the opposite is true. **A closer look at the specific items indicates that the items *were happy* and *couldn't get going* were most at risk for additive response bias, with their intercepts being variant in one third of the groups. The items *felt lonely* and *felt sad* showed the most stable conditions across the groups, and were thus least a risk for additive response bias.**

In sum, the findings for all models in Table 1 indicate that the one-dimensional CES-D 8 scale with correlated errors of the reverse-worded items showed configural, metric, and partial scalar invariance across all countries and gender groups in the analysis. At this level of invariance, comparison of latent (co)variances **and latent** means of the CES-D 8 across gender and countries, is warranted.

Comparison of the Observed and Latent Means

The previously mentioned results suggest that the CES-D 8 measures the same construct both in gender and in countries included in the analysis. **Based on the partial metric invariance model, we estimated the latent means of the CES-D 8 for men and women in each country separately. These latent means can be regarded as very conservative, gender- and culture-neutral estimates of gender differences in depression. Since identification of the model requires the factor loading of the item *felt depressed* to be set to 1 and its intercept to 0, the scale of the latent means is arbitrary (Meuleman, Davidov, and Billiet 2009; Gregorich2006). Therefore, interpretation of the absolute values is not useful, but comparisons of the rankings in the overall depression level and gender gap are more informative.** Results are shown in Table 2, along with the observed means and their standard deviations.

Insert Table 2 here

The observed means indicate that overall depression rates (results not shown) are clustered together by region, with the highest scores in Eastern and Central European countries, and the lowest scores in Western and Northern European countries. The Norwegian population reports the lowest CES-D 8 scores, followed by Denmark and Switzerland, while the highest mean scores were found in the Ukraine, Hungary, and the Russian Federation. However gender differences in depression do not show a similar trend. The former Soviet countries do show higher gender differences than the other countries; this is also the case in the Southern European countries. However, in all countries but Ireland and Finland, female respondents score significantly higher on the CES-D 8 scale than male respondents. The gender difference is largest in Portugal ($\Delta = 1.83$, $p < 0.000$) and smallest in Ireland ($\Delta = 0.10$, n.s.). Ukrainian females and Hungarian males report the highest depression level of their sex, with a mean score of, respectively, 17.44 and 16.13. Lowest depression levels in each sex are reported by Norwegian females (mean of 12.49) and Norwegian males (mean of 12.04).

The impact of our model at the level of partial metric invariance does not modify the ranking of the countries by overall mean depression score much. The latent means are highest in the Eastern and Central European countries, and lowest in Northern and Western European countries. Gender differences in depression are confirmed, with significantly higher scores for females than males in all countries but Ireland. The highest gender differences in depression are confirmed by the **latent means** in the former Soviet Union, Poland, and in the Southern European countries, with a substantial increase in the latter. The **latent means** indicate the largest difference in depression in Portugal (Δ = **0.33, p<0.000**), followed by Cyprus (Δ = **0.30, p<0.000**). The smallest differences were found in Ireland (Δ = **0.02**, n.s.) and Latvia (Δ = **0.05, p<0.05**).

Our observed means overestimated the level of depression most in the German and Finnish population, and underestimated it the most in the Spanish and Polish population, compared to the other countries. The impact of our model on the cross-national ranking of gender differences in depression is highest in Latvia, followed by Slovenia and Finland. In the former the estimated gender difference is much lower than the observed difference compared to the other countries; in the latter two the opposite is true.

These differences in the ranking of the countries based on the observed depression means and the **latent means** signal the impact of gender- and culture-related symptom differences in depression. In general, depression is most strongly related to *feeling depressed* and is least strongly related to *lack of enjoyment in life*. However, somatic complaints carry more weight in most of the Central and Eastern European countries, while in the remaining countries this is the case only in Austrian and Cypriot men. In the Northern European countries mood affects are more pivotal than somatic complaints. Finally, the symptoms *felt sad*, *effort,* and *get going* showed the largest gender*country variation in expression.

Discussion

Women generally report more complaints of depression. Cross-national comparative research and meta-analyses usually put the gender ratio in the prevalence of depression in Western countries at approximately 2:1, although the same studies report substantial cross-national and cross-cultural variation (Kessler et al. 1993; Nolen-hoeksema 1990; Weissman et al. 1984), even within the more homogeneous sample of Western advanced economies. This variability points to important societallevel determinants of mental ill health in women and men. Nevertheless, comparative research explaining this variability is hampered by a lack of useable data. The third round of the European Social Survey, utilized in our research, uses the same research design and the same depression inventory across a substantial set of European countries. It includes a shorter version of the CES-D, allowing for the most comprehensive estimation of the frequency and severity of symptoms of depression in women and men across Europe, and, so far, covers 25,490 women and 21,179 men in 25 European countries.

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 made use of MCFA in order to establish measurement invariance of the CES-D 8 across gender and countries. A one-dimensional depression model, with all items loading on the factor depression and with correlated errors between the reverse-worded items *were happy* and ***enjoyed life***, fit the data best. Measurement invariance was established at the level of configural, scalar, and partial metric invariance. **Partial metric invariance was obtained by relaxing 83 intercepts. The items *were happy* and *couldn't get going* were most at risk for additive response bias, while the items *felt lonely* and *felt sad* were least at risk. In addition the Belgian, Estonian, Irish and Slovenian populations show very stable conditions, while the CES-D8 item scores of the German males and French females were most invariant.** Our results indicate that the CES-D 8 scale can be used to compare (co)variances, observed means, and latent means in depression of men and women across Europe.

Next we estimated the latent means of de CES-D 8 across gender and countries, eliminating all measurement artifacts. To the best of our knowledge, this study is the first to present information on gender differences in the frequency and severity of symptoms of depression in 25 European countries using a depression inventory freed from extraneous influences on observed differences in responses that are unrelated to actual levels of depression. Our study confirms the consistent epidemiological finding that women report more complaints of depression than men, although there is substantial cross-national variation. First, as concerns the gender ratio, it is important to note that in both the Irish and the Finnish sample no gender differences in observed mean depression scores were found. However after correcting for the partial metric model, only the gender difference in the Irish population stays insignificant. The **latent means** also indicate highest gender differences in the Southern European countries, the former Soviet countries, and Poland. Comparing our results with those of other cross-national comparisons based on survey data (Hopcroft and Bradley 2007) or on meta-analyses of the results of surveys (Immerman and Mackey 2003) makes clear that it is difficult to consistently rank countries according to the gender gap in depression across studies.

While the average gender difference points to more universal genetic, neuro-hormonal or psychobiological gender-linked causes of depression (Kuehner2003), its cross-national variation suggests that social conditions have an important impact too (Weissman et al.1996). Social models to explain gender differences in depression have emphasized the activities and circumstances of women's and men's everyday lives as sources of stress. Both within and outside the family, female roles seems more prone to role limitations associated with lack of choice, to role overload, to competing social roles and to role underevaluation (Piccinelli and Wilkinson2000). However, the impact of these determinants might be modified by macro-level conditions, such as gendered welfare state regimes and levels of de-familialisation (Bambra et al.2008;Bambra2007), gender stratification systems (Hopcroft and Bradley2007;Immerman and Mackey2003), but also gender beliefs and ideologies (Chafetz1990;Stoppard2000). However, these causes for cross-national variation in the gender ratio in depression have seldom been empirically tested. In subsequent analyses of the present data, we will explore these and other possible macro-sociological determinants.

Some limitations of our study are worth noting when interpreting the results. When testing measurement invariance in large community samples such as the ESS-3, the researcher should bear in mind that the variables of interest are often non-normally distributed, specifically when working with ordinal Likert scales (Lubke and Muthen 2004). 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 that compensates for the normality assumption (Nevitt and Hancock 1997). 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 estimations methods even after controlling for non-normality.

In our analysis some intercepts were found to vary across our groups, whereas other intercepts were invariant, resulting in partial metric invariance. We based our decision to relax a certain intercept on the MIs and EPCs as provided by AMOS 16.0. This was done only if the intercept to be relaxed could be substantially interpreted. How should we then interpret our partial metric model? In our analysis, 85 of the 400 intercepts were relaxed. There are three options for what to do with our scale. One option might be to omit the items that were found to perform differently across groups. A second option might be to retain all 8 items of the CES-D 8 scale in the belief that the population differences in factor structure are "small" in some sense and that these differences will not obscure inferences from the scale. A third option might be to abandon the use of the scale altogether for comparisons across populations, reasoning that the lack of invariance establishes that the scale is measuring different latent variables in the two populations. Unfortunately, the current literature on factorial invariance offers little guidance for choosing among these three options (Millsap and Kwok2004).

It is important to note that we did not take into account differences between countries in either demographic or socioeconomic characteristics of the populations—i.e., the age structure of the population, divorce, and unemployment rates—or gendered welfare state regimes (Bambra et al. 2008) or gender stratification systems (Hopcroft and Bradley 2007; Immerman and Mackey 2003) as possible causes for cross-national variation in the gender ratio in depression. In subsequent analyses of the present data, we will explore these and other possible macro-sociological determinants.

Finally, the current findings do not automatically imply psychometric equivalence across social groups distinguished by other criteria such as language, ethnicity, social class, or age. All of these social groups may have group-specific attributes that lead to measurement inequivalence of (self-report) scales. We therefore strongly suggest testing measurement invariance before comparing specific group scores. 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.

Reference List

Alonso, J., M. C. Angermeyer, S. Bernert, R. Bruffaerts, I. S. Brugha, H. Bryson, G. de Girolamo, R. De Graaf, K. Demyttenaere, I. Gasquet, J. M. Haro, S. J. Katz, R. C. Kessler, V. Kovess, J. R. Lepine, J. Ormel, G. Polidori, L. J. Russo, G. Vilagut, J. Almansa, S. rbabzadeh-Bouchez, J. Autonell, M. Bernal, M. A. Buist-Bouwman, M. Codony, A. Domingo-Salvany, M. Ferrer, S. S. Joo, M. Martinez-Alonso, H. Matschinger, F. Mazzi, Z. Morgan, R. Morosini, C. Palacin, B. Romera, N. Taub, and W. A. M. Vollebergh. 2004. "Prevalence of Mental Disorders in Europe: Results From the European Study of the Epidemiology of Mental Disorders (ESEMeD) Project." *Acta Psychiatrica Scandinavica* 109:21-27.

Angst, J., A. Gamma, M. Gastpar, J. P. Lepine, J. Mendlewicz, and A. Tylee. 2002. "Gender Differences in Depression - Epidemiological Findings From the European DEPRES I and II Studies." *European Archives of Psychiatry and Clinical Neuroscience* 252(5):201-9.

Arbuckle, J. L. 2007. *AMOS 16.0 User's Guide.* Chicago: SPSS Inc.

Ayuso-Mateos, J. L., J. L. Vázquez-Barquero, C. Dowrick, V. Lehtinen, C. Wilkinson, L. Lasa, H. Page, G. Dunn, and G. Wilkinson. 2001. "Depressive Disorders in Europe: Prevalence Figures From the ODIN Study." *British Journal of Psychiatry* 179:308-16.

Bambra, C. 2007. "Defamilisation and Welfare State Regimes: a Cluster Analysis." *International Journal of Social Welfare* 16(4):326-38.

Bambra, Clare, Daniel P. Pope, Viren Swami, Debbi L. Stanistreet, Albert Jan Roskam, A. E. Kunst, and Alex Scott-Samuel. 3-9-2008. "Gender, Health Inequalities and Welfare State Regimes: a Cross-National Study of Thirteen European Countries." *Journal of Epidemiology and Community Health* doi:10.1136/jech.2007.070292.

Baumgartner, H. and J. B. E. M. Steenkamp. 2001. "Response Styles in Marketing Research: A Cross-National Investigation." *Journal of Marketing Research* 38(2):143-56.

Bebbington, P. 1996. "The Origins of Sex Differences in Depressive Disorder: Bridging the Gap." *International Review of Psychiatry* 8(4):295-332.

Bentler, P. M. 1990. "Comparative Fit Indexes in Structural Models." *Psychological Bulletin* 107:238-46.

Berkman, L. F., C. S. Berkman, S. Kasl, D. H. Freeman, L. Leo, A. M. Ortfeld, Connonihuntley J., and J. A. Brody. 1986. "Depressive Symptoms in Remation to Physical Health and Functioning in the Elderly." *American Journal of Epidemiology* 124:372-88.

Billiet, J. and M. McClendon. 2000. "Modeling Acquiescence in Measurement Models for Two Balanced Sets of Items." *Structural Equation Modeling* 7(4):608-28.

Bollen, K. A. 1989. *Structural Equations With Latent Variables.* New York: Wiley.

Börsch-Supan, A., A. Brugiavini, H. Jürges, J. Mackenbach, J. Siegrist, and G. Weber, Börsch-Supan, A., A. Brugiavini, H. Jürges, J. Mackenbach, J. Siegrist, and G. Weber. 2005. *Health, Ageing and Retirement in Europe – First Results from the Survey of Health, Ageing and Retirement in Europe.* MEA: University of Mannheim.

Bouma, J., A. V. Ranchor, R. Sanderman, and E. Ronderson, Bouma, J., A. V. Ranchor, R. Sanderman, and E. Ronderson. 1995. *"Symptomen van depressie (CES-D). Het meten van symptomen van depressie met de CES-D, een handleiding."* (Symptoms of depression (CES-D). Measuring the symptoms of depression with the CES-D, a guide). Retrieved November 3, 2008 (http://www.rug.nl/gradschoolshare/research_tools/ assessment_tools/CES-D_handleiding.pdf)

Brown, T. A. 2006. *Confirmatory Factor Analysis in Applied Research.* New York: The Guildford Press.

Browne, M. W. and R. Cudeck. 1992. "Alternative Ways of Assessing Model Fit." *Sociological Methods and Research* 21:230-258.

Byrne, B. M. 1989. "Multigroup Comparisons and the Assumption of Equivalent Construct Validity Across Groups: Methodological and Substantive Issues." *Multivariate Behavioral Research* 24:503-23.

Byrne, B. M., P. Baron, and T. L. Campbell. 1993. "Measuring Adolescent Depression: Factorial Validity and Invariance of the Beck Depression Inventory Across Gender." *Journal of Research on Adolescence* 3:127-43.

Byrne, B. M., R. J. Shavelson, and B. Muthén. 1989. "Testing for the Equivalence of Factor Covariance and Mean Structures: The Issue of Partial Measurement Invariance." *Psychological Bulletin* 105:456-66.

Callahan, C. M. and F. D. Wolinsky. 1994. "The Effect of Gender and Race on the Measurement Properties of the CES-D in Older Adults." *Medical Care* 32(4):341-56.

Chafetz, J. S. 1990. *Gender Equity. An Integrated Theory of Stability and Change.* Newbury Park, California: Sage Publications, Inc.

Cheung, G. W. and R. B. Rensvold. 2000. "Assessing Extreme and Acquiescence Response Sets in Cross-Cultural Research Using Structural Equations Modeling." *Journal of Cross-Cultural Psychology* 31:187-212.

Clark, V. A., C. S. Aneshensel, R. R. Frerichs, and T. M. Morgan. 1981. "Analysis of Effects of Sex and Age in Response to Items on the CES-D Scale." *Psychiatry Research* 5(2):171-81.

Cohen, J., P. Cohen, West S.G., and L. S. Aiken. 2003. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* Mahwah, New Jersey: Erlbaum.

Cole, S. R., I. Kawachi, Maller S.J., and L. F. Berkman. 2000. "Test of Item-Response Bias in the CES-D Scale: Experience From the New Haven EPESE Study." *Journal of Clinical Epidemiology* 53:285-89.

Fuhrer, R. and F. Rouillon. 1989. "La Version Française De L'Échelle CES-D (Center for Epidemiologic Studies-Depression Scale). Description Et Traduction De L'Échelle D'Autoévaluation." (The French version of the CES-D scale. Description and translation of a self-report instrument.) *Psychiatrie Psychobiologie* 4(3):163-166.

Golding, J. M. and C. S. Aneshensel. 1989. "Factor Structure of the Center of Epidemiologic Studies Depression Scale Among Mexican Americans and Non-Hispanic Whites." *Journal of Consulting and Clinical Psychology* 1:163-68.

Goncalves, B. and T. Fagulha. 2004. "The Portuguese Version of the Center for Epidemiologic Studies Depression Scale (CES-D)." *European Journal of Psychological Assessment* 20(4):339-48.

Gregorich, S. E. 2006. "Do Self-Report Instruments Allow Meaningful Comparisons Across Diverse Population Groups? Testing Measurement Invariance Using the Confirmatory Factor Analysis Framework." *Medical Care* 44:S78-S94.

Hautzinger, M. 1988. "Die CES-D Skala. Ein Depressionsinstrument Für Untersuchungen in Der Allgemeinbevölkerung." (The CES-D Scale. A depression instrument for general population research). *Diagnostica 34*.

Hertzog C, Van Alstine J, Usala PD, Hultsch DF, and Dixon R. 1990. "Measurement Properties of the Center for Epidemiological Studies Depression Scale (CES-D) in Older Populations." *Psychological Assessment* 2(1):64-72.

Hopcroft, R. L. and D. B. Bradley. 2007. "The Sex Difference in Depression Across 29 Countries." *Social Forces* 85(4):1483-507.

Horn, J. L. and J. McArdle. 1992. "A Practical and Theoretical Guide to Measurement Invariance in Aging Research." *Experimental Aging Research* 18:117-44.

Hu, L. T. and P. M. Bentler. 1998. "Fit Indices in Covariance Structure Modeling: Sensitivity to Underparameterized Model Misspecification." *Psychological Methods* 3(4):424-53.

Immerman, R. S. and W. C. Mackey. 2003. "The Depression Gender Gap: A View Through a Biocultural Filter." *Genetic Social and General Psychology Monographs* 129(1):5-39.

Joseph, S. and C. A. Lewis. 1995. "Factor-Analysis of the Center for Epidemiologic Studies-Depression Scale." *Psychological Reports* 76(1):40-42.

Jowell, R., 2007. *European Social Survey 2006/2007. Round 3: Technical Report*. London: Citiy University, Centre for Comparative Social Surveys.

Kessler, R. C., K. A. Mcgonagle, M. Swartz, D. G. Blazer, and C. B. Nelson. 1993. "Sex and Depression in the National Comorbidity Survey .1. Lifetime Prevalence, Chronicity and Recurrence." *Journal of Affective Disorders* 29(2-3):85-96.

Kuehner, C. 2003. "Gender Differences in Unipolar Depression: an Update of Epidemiological Findings and Possible Explanations." *Acta Psychiatrica Scandinavica* 108(3):163-74.

Lepine, J. P., M. Gastpar, J. Mendlewicz, and A. Tylee. 1997. "Depression in the Community: The First Pan-European Study DEPRES (Depression Research in European Society)." *International Clinical Psychopharmacology* 12(1):19-29.

Lubke, G. H. and B. O. Muthen. 2004. "Applying Multigroup Confirmatory Factor Models for Continuous Outcomes to Likert Scale Data Complicates Meaningful Group Comparisons." *Structural Equation Modeling* 11:514-34.

Maier, W., M. Gansicke, R. Gater, M. Rezaki, B. Tiemens, and R. F. Urzua. 1999. "Gender Differences in the Prevalence of Depression: a Survey in Primary Care." *Journal of Affective Disorders* 53(3):241-52.

Marsch, H. W. 1996. "Positive and Negative Global Self-Esteem: A Substantively Meaningful Distinction or Artifactors?" *Journal of Personality and Social Psychology* 70:810-819.

Meads, D. M., S. P. McKenna, and L. C. Doward. 2006. "Assessing the Cross-Cultural Comparability of the Centre for Epidemiologic Studies Depression Scale (CES-D)." Poster presented at the

International Society for Pharmacoeconomics and Outcome Research, 2006. Retrieved November 3, 2008 (http://www.ispor.org/awards/9euro/MeadsPMH46.pdf)

Meredith, W. 1993. "Measurement Invariance, Factor-Analysis and Factorial Invariance." *Psychometrika* 58:525-43.

Meuleman, B., E. Davidov, and J. Billiet. 2009. "Changing Attitudes Toward Immigration in Europe, 2002-2007: A Dynamic Group Conflict Theory Approach." *Social Science Research* 38(2):352-65.

Millsap, R. E. and O. M. Kwok. 2004. "Evaluating the Impact of Partial Factorial Invariance on Selection in Two Populations." *Psychological Methods* 9(1):93-115.

Mirowsky, J. J. and C. E. Ross. 1995. "Sex Differences in Distress - Real or Artifact." *American Sociological Review* 60(449):468.

Moors, G. 2004. "Facts and Artifacts in the Comparison of Attitudes Among Ethnic Minorities. A Multigroup Latent Class Structure Model With Adjustment for Response Style Behavior." *European Sociological Review* 20(4):303-20.

Nevitt, J. and G. R. Hancock. 1997. “Relative Performance of Rescaling and Resampling Approaches to Model Chi-Square and Parameter Standard Error Estimation in Structural Equation Modeling.” Paper presented at the annual meeting of the American Educational Research Association, San Diego, April 14, 1997.

Nolen-hoeksema, S. 1990. *Sex Differences in Depression* Stanford: Stanford University Press.

Perreira, K. M., N. Deeb-Sossa, K. M. Harris, and K. Bollen. 2005. "What Are We Measuring? An Evaluation of the CES-D Across Race/Ethnicity and Immigrant Generation." *Social Forces* 83(4):1567-601.

Piccinelli, M. and G. Wilkinson. 2000. "Gender Differences in Depression - Critical Review." *British Journal of Psychiatry* 177:486-92.

Radloff, L. S. 1977. "The CES-D Scale: A Self-Report Depression Scale for Research in the General Population." *Applied Psychological Measurement* 1:385-401.

Riddle, A. S. and U. Hess, Riddle, A. S. and U. Hess. 2008. “Static versus dynamic structural models of depression: The case of the CES-D.” Retrieved November 3, 2008 (http://www.cirano.qc.ca/pdf/publication/2002s-37.pdf)

Riordan, C. M. and R. J. Vandenberg. 1994. "A Central Question in Cross-Cultural Research - Do Employees of Different Cultures Interpret Work-Related Measures in An Equivalent Manner." *Journal of Management* 20(3):643-71.

Rorer, L. G. 1965. "The Great Response-Style Myth." *Psychological Bulletin* 63:123-56.

Shafer, A. B. 2006. "Meta-Analysis of the Factor Structures of Four Depression Questionnaires: Beck, CES-D, Hamilton, and Zung." *Journal of Clinical Psychology* 62(1):123-46.

Spijker, J., F. B. van der Wurff, E. C. Poort, C. H. M. Smits, A. P. Verhoeff, and A. T. F. Beekman. 2004. "Depression in First Generation Labour Migrants in Western Europe: the Utility of the Center for Epidemiologic Studies Depression Scale (CES-D)." *International Journal of Geriatric Psychiatry* 19(6):538-44.

Steenkamp, J. B. E. M. and H. Baumgartner. 1998. "Assessing Measurement Invariance in Cross-National Consumer Research." *Journal of Consumer Research* 25:78-90.

Steffick, D., Steffick, D. 2000. *Documentations of affective functioning measures in the Health and Retirement Study. HRS/AHEAD Documentation Report DR-005.* Ann Arbor: Survey Research Center, University of Michigan.

Steiger, J. H. 1990. "Structural Model Evaluation and Modification – an Interval Estimation Approach." *Multivariate Behavioral Research* 25:173-80.

Stommel, M., B. A. Given, C. W. Given, H. A. Kalaian, R. Schulz, and R. Mccorkle. 1993. "Gender Bias in the Measurement Properties of the Center for Epidemiologic Studies-Depression Scale (CES-D)." *Psychiatry Research* 49(3):239-50.

Stoppard, J. M. 2000. *Understanding Depression. Feminist Social Constructionist Approaches* New York: Routledge.

Tucker, L. R. and C. Lewis. 1973. "Reliability Coefficients for Maximum Likelihood Factor-Analysis." *Psychometrika* 38:1-10.

Vandenberg, R. J. and C. E. Lance. 2000. "A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research." *Organisation Research Methods* 2:4-69.

vandeVijver, F. J. R. 2003. "Cross-Cultural Survey Methods." Pp. 143-155 in *Bias and Equivalence,* edited by J.Harkness, F.Van de Vijver, and P.Mohler. New Jersey: Wiley & Sons.

vandeVijver, F. J. R. and Y. H. Poortinga. 1997. "Towards an Integrated Analysis of Bias in Cross-Cultural Assessment." *European Journal of Psychological Assessment* 13(1):29-37.

Vega, W. A. and R. G. Rumbaut. 1991. "Ethnic Minorities and Mental Health." *Annual Review of Sociology* 17:351-83.

Wauterickx, N. and P. Bracke. 2005. "Unipolar Depression in the Belgian Population - Trends and Sex Differences in an Eight-Wave Sample." *Social Psychiatry and Psychiatric Epidemiology* 40(9):691-99.

Weissman, M. M., P. J. Leaf, C. E. Holzer, J. K. Myers, and G. L. Tischler. 1984. "The Epidemiology of Depression - An Update on Sex-Differences in Rates." *Journal of Affective Disorders* 7(3-4):179-88.

Weissman, M. M., R. C. Bland, G. J. Canino, C. Faravelli, S. Greenwald, H. G. Hwu, P. R. Joyce, E. G. Karam, C. K. Lee, J. Lellouch, J. P. Lepine, S. C. Newman, M. RubioStipec, J. E. Wells, P. J. Wickramaratne, H. U. Wittchen, and E. K. Yeh. 1996. "Cross-National Epidemiology of Major Depression and Bipolar Disorder." *Jama-Journal of the American Medical Association* 276(4):293-99.

World Health Organization (WHO) (2000). *Women's Mental Health. An Evidence Based Review.* Retrieved November 3,2008 (http://whqlibdoc.who.int/hq/2000/WHO_MSD_MDP_00.1.pdf)

Table 1

Model Fit Summary: chi-square, CFI, TLI and RMSEA. ESS-3, 2006–2007 (Jowell, 2007)[6]

Model	χ^2	df	Sign.	CFI	TLI	RMSEA
Best fitting model						
1a. One dimensional	12793.362	20	0.000	0.894	0.852	0.117
1b. Two dimensional	11585.181	19	0.000	0.904	0.858	0.114
1c. One dimensional - correlated errors	2715.030	19	0.000	0.978	0.967	0.055
1d. Two dimensional - correlated errors	1839.087	18	0.000	0.985	0.976	0.047
MCFA Equivalence tests						
2. Configural	5026.743	950	0.000	0.964	0.946	0.010
3a. Metric	6895.473	1293	0.000	0.950	0.946	0.010
4a. Scalar	15659.113	1636	0.000	0.875	0.893	0.014
4b. Partial Scalar	**9580.847**	**1553**	**0.000**	**0.928**	**0.935**	**0.011**

[6] **The TuckerLewis index (TLI) and the Comparative Fit index (CFI) provide evidence for a good fit if 0.90 or higher. The Root Mean Squared Error of Approximation (RMSEA) indicates a good fit when the score is 0.05 or less.**

Table 2.

Comparison of Observed Means and Standard Deviations with Latent Means and Standard Deviations. ESS-3, 2006–2007 (Jowell, 2007)

	OBSERVED MEANS					LATENT MEANS					RANKING Observed → Latent	
	Male		Female		Δ; p	Male		Female		Δ; p	Country Mean	Gender difference
	Mean	S.D.	Mean	S.D.		Mean	S.D.	Mean	S.D.			
Austria	13.23	3.77	13.75	4.16	0.53; 0.001	1.34	0.41	1.49	0.41	0.14; 0.000	10 →10	5 → 8
Belgium	12.74	3.77	14.07	4.34	1.33; 0.000	1.32	0.41	1.56	0.41	0.23; 0.000	9 → 11	18 → 16
Bulgaria	15.25	4.79	16.54	5.02	1.29; 0.000	1.71	0.60	1.91	0.60	0.20; 0.000	22 → 22	17 → 15
Switzerland	12.39	3.14	13.08	3.51	0.69; 0.000	1.17	0.31	1.34	0.31	0.18; 0.000	3 → 4	7 → 11
Cyprus	12.47	3.14	14.02	3.90	1.57; 0.000	1.11	0.29	1.41	0.29	0.30; 0.000	7 → 5	22 → 24
Germany	13.65	3.47	14.49	3.92	0.84; 0.000	1.32	0.37	1.51	0.37	0.19; 0.000	15 → 9	10 → 12
Denmark	12.50	3.12	13.02	3.51	0.53; 0.002	1.12	0.31	1.25	0.30	0.13; 0.000	2 → 1	6 → 6
Estonia	14.32	3.77	15.16	4.14	0.85; 0.000	1.46	0.41	1.63	0.41	0.17; 0.000	17 → 16	11 → 10
Spain	12.86	3.81	14.35	4.53	1.49; 0.000	1.40	0.45	1.67	0.45	0.27; 0.000	12 → 15	21 → 21
Finland	12.82	3.14	13.11	3.48	0.29; 0.057	1.13	0.30	1.26	0.30	0.13; 0.000	5 → 2	2 → 7
France	12.90	3.74	14.25	4.63	1.35; 0.000	1.31	0.41	1.59	0.41	0.28; 0.000	11 → 12	19 → 23
United Kingdom	13.44	4.03	14.16	4.30	0.72; 0.000	1.40	0.43	1.52	0.43	0.13; 0.000	14 → 13	8 → 5
Hungary	16.13	4.98	17.08	5.21	0.95; 0.000	1.91	0.63	2.06	0.63	0.15; 0.000	24 → 25	13 → 9
Ireland	12.84	3.67	12.93	3.61	0.10; 0.579	1.28	0.40	1.31	0.40	0.02; 0.230	4 → 6	1 → 1
Latvia	15.48	3.85	16.21	3.77	0.73; 0.000	1.68	0.45	1.72	0.45	0.05; 0.018	21 → 20	9 → 2
Netherlands	12.71	3.50	13.88	3.95	1.17; 0.000	1.28	0.37	1.47	0.37	0.19; 0.000	8 → 8	16 → 14
Norway	12.04	3.01	12.49	3.22	0.46; 0.002	1.16	0.32	1.25	0.32	0.09; 0.000	1 → 3	3 → 3
Poland	13.92	4.43	15.29	5.15	1.36; 0.000	1.52	0.51	1.78	0.51	0.26; 0.000	16 → 19	20 → 20
Portugal	14.64	3.95	16.47	4.74	1.83; 0.000	1.56	0.47	1.88	0.47	0.33; 0.000	20 → 21	25 → 25
Romania	14.78	3.83	15.92	4.02	1.14; 0.000	1.46	0.42	1.65	0.42	0.19; 0.000	18 → 17	15 → 13
Russian Fed.	15.12	4.35	16.83	4.67	1.71; 0.000	1.70	0.52	1.95	0.52	0.26; 0.000	23 → 23	23 → 19
Sweden	12.46	3.41	13.52	4.15	1.06; 0.000	1.17	0.34	1.42	0.34	0.25; 0.000	6 → 7	14 → 18
Slovenia	13.33	3.28	14.21	4.23	0.88; 0.000	1.33	0.34	1.58	0.34	0.25; 0.000	13 → 14	12 → 17
Slovakia	15.18	3.82	15.69	4.08	0.51; 0.007	1.53	0.41	1.65	0.41	0.13: 0.000	19 → 18	4 → 4
Ukraine	15.69	4.67	17.44	5.07	1.75; 0.000	1.73	0.55	2.00	0.55	0.27; 0.000	25 → 24	24 → 22
Total	13.67	3.96	14.81	4.50	1.14; 0.000	1.40	0.47	1.61	0.55	0.22; 0.000		