Psychological Assessment

Testing Measurement Invariance of the Depression, Anxiety, and Stress Scales (DASS-21) Across Four Countries
Saskia Scholten, Julia Velten, Angela Bieda, Xiao Chi Zhang, and Jürgen Margraf

CITATION
http://dx.doi.org/10.1037/pas0000440
Testing Measurement Invariance of the Depression, Anxiety, and Stress Scales (DASS-21) Across Four Countries

Saskia Scholten, Julia Velten, Angela Bieda, Xiao Chi Zhang, and Jürgen Margraf
Ruhr-Universität Bochum

The rising burden of mental and behavioral disorders has become a global challenge (Murray et al., 2012). Measurement invariant clinical instruments are necessary for the assessment of relevant symptoms across countries. The present study tested the measurement invariance of the 21-item version of the Depression, Anxiety, and Stress Scales (DASS; Lovibond & Lovibond, 1995b) in Poland, Russia, the United Kingdom (U.K.), and the United States of America (U.S.). Telephone interviews were conducted with population-based samples ($n_{PL} = 1003$, $n_{RU} = 3020$, $n_{U.K.} = 1002$, $n_{U.S.} = 1002$). The DASS-21 shows threshold measurement invariance. Comparisons of latent means did not indicate differences between U.K. and U.S. samples. However, Polish and Russian samples reported more depressive symptoms compared with U.K. and U.S. samples; the Russian sample had the highest levels of anxiety symptoms and the Polish sample demonstrated the highest stress levels. The DASS-21 can be recommended to meaningfully compare the relationships between variables across groups and to compare latent means in Polish-, Russian-, and English-speaking populations.

Public Significance Statement
This study underlines the importance to test equivalence and examine potential bias in cross-cultural research frameworks. We successfully tested the cross-cultural applicability of a questionnaire for symptoms of depression, anxiety, and stress and found that the populations in Poland and Russia reported more symptoms of depression, anxiety, and stress than the populations in the U.K. and the U.S.

Keywords: cross-cultural comparison, Depression, Anxiety and Stress Scales, measurement invariance, psychological assessment

Supplemental materials: http://dx.doi.org/10.1037/pas0000440.supp

Across countries the increase of anxiety and mood disorders has become a widespread challenge. Worldwide, anxiety disorders are found most often with a 12-month prevalence of 11%. Although the average 12-month prevalence of mood disorders is slightly lower (6%), mood disorders are still the second largest group of mental disorders worldwide (Kessler et al., 2009). The global impact of depressive and anxiety disorders becomes especially salient in terms of disability adjusted life years (DALY). Depressive disorders accounted for 41% and anxiety disorders accounted for 15% of DALY caused by mental and substance use disorders in 2010 (Murray et al., 2012).

However, prevalence rates vary worldwide: In Europe, 38% of the population suffers from a mental disorder each year. Anxiety disorders take up the greatest part with 14%. Major depression accounts for 7% of the disorders (Wittchen et al., 2011). In the United States the picture is similar; 12-month prevalence of anxiety disorders is 18%, followed by mood disorders with 10% (Kessler, Chiu, Demler, & Walters, 2005). But in Central Asia estimated point prevalence for anxiety disorders is 4% for men and 6% in women (Baxter et al., 2014) whereas symptoms of anxiety were much higher in a population-based survey in Arkhangelsk in Russia (22% of the men and 53% of the women), again followed by symptoms of depression (11% of the men and 34% of the women) (Averina et al., 2005). In a cross-sectional study of urban population samples in Russia, Poland, and the Czech Republic, the point-prevalence of depressive symptoms was around 20% in men and 40% in women (Bobak et al., 2006).

Differences in prevalence rates do not necessarily represent real cross-cultural differences. Specific symptoms might belong to the...
range of culturally normative experiences in one culture and might be actual signs of distress in another (Kirmayer, 2001; López & Guarnaccia, 2000). So called culture-bound syndromes describe a cluster of symptoms that only appear in a specific cultural context (Weiss & Somma, 2007). Yet, even if the same symptom cluster appears across cultures, methodological issues can still lead to false conclusions about cross-cultural differences (Bowden & Fox-Rushby, 2003; van de Vijver & Leung, 1997; van de Vijver & Tanzer, 2004). For example, the description of a certain disorder might differ across countries, making a word-by-word translation unfeasible to assess symptoms of the specific disorder (Ryder et al., 2008); or systematic bias in the assessment itself might appear as a result of answer tendencies or familiarity with the question format (Byrne, 2016). More than 50 years of research have addressed these problems (Poortinga, 1995), but best-practice strategies of cross-cultural assessment and comparisons are just recently entering mainstream psychological research (Byrne, 2016).

The purpose of the present study was to test the cross-cultural applicability of the 21-item version of the Depression, Anxiety, and Stress Scales (DASS-21; Lovibond & Lovibond, 1995a) in two Western and two Eastern European countries and to compare the means of symptoms of depression, anxiety, and stress across those countries if applicability is given. In the introduction, we therefore first outline central issues of cross-cultural assessment. Then, we provide an overview of the DASS-21 and its applicability within and across cultures today. Finally, we describe the aim of the study in detail.

Cross-Cultural Assessment

The earliest approach to cross-cultural psychology was transport and test (Berry, 1969; Berry, Poortinga, Segall, & Dasen, 2002). Here, tests and interventions established in Europe or the U.S. were literally exported and imposed on local cultures without cultural adaptation (F. M. Cheung, 2012). Without further reflection, the Euro/American-centered view was assumed to be universal. Today, cross-cultural differences are taken into account as important sources of information to understand psychosocial mechanisms of human behavior, interaction, and society. To avoid former culture-blind mistakes, it is important that research approaches are equally valid across cultures. van de Vijver and Leung (1997) distinguish between three levels of equivalence: construct equivalence, measurement equivalence, and scalar equivalence. With respect to construct equivalence, the same constructs are under study across different cultural groups even though they might be assessed differently. If metric measures have the same measurement unit, measurement equivalence is given. Finally, scalar equivalence describes a case where not only the measurement unit is the same, but also the origin or scale.

Within cross-cultural psychology, special consideration needs to be given to cross-cultural assessment. Still too often, it is assumed that a translated instrument measures the same phenomena in the same way in a different culture (Byrne, 2016). However, this assumption needs to be tested and if it does not hold, the measurement instrument should be adapted accordingly. Most frequently assessment instruments are only adopted from one culture to another which refers to a close translation of an instrument into a target language (He & van de Vijver, 2012). The procedure is valid if the construct and the instrument features are suitable in all cultural groups involved. Yet, this condition is rarely fulfilled. Adaptation of assessment instruments considers this problem and expands the mere translation as a linguistic task through psychometric evaluation of the instrument in the new context (He & van de Vijver, 2012). An example of a detailed description of test adaptation procedures is provided in the guidelines of the International Test Commission (Hambleton, 2005; International Test Commission, 2005, 2016). If adoption and adaptation do not yield satisfactory results, a new instrument needs to be compiled which is called assembly. The new instrument should be maximally culturally suitable, however, throughout the process the comparability across cultures decreases (He & van de Vijver, 2012).

For the present study, cross-cultural adaptation is the relevant approach because we aimed to test the cross-cultural applicability of the DASS-21. First, both the instrument and the measured construct need to be rigorously validated for the conversion of an assessment scale from one language and cultural application for the use in another language and culture. Then, the levels of equivalence mentioned above need to be tested for the respective assessment instrument (Byrne, 2016). In this more specific context, the term measurement invariance is used synonymously for measurement equivalence to describe equivalence as a property of a measurement instrument (Davidov, Meuleman, Cicciuč, Schmidt, & Billiet, 2014). To evaluate whether a translated version is comparable across countries, commonly three successive more restricted levels of invariance or equivalence are tested (Byrne, 2008, 2016; Hirschfeld & Von Brachel, 2014; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000): On the first level configural invariance is tested. It implies that the configuration of the model, the number of factors and their loading patterns, are the same across groups without imposing further equality constraints on the model parameters. On the second level, all factor loadings are constrained to be equal across groups to test metric invariance. It tests to what extent the items designed to measure the same factor are equivalent across groups. If the assumption holds, the factorial validity of an instrument is equivalent across samples. In contrast, latent means can only be meaningfully compared across different groups if, in addition to the factor loadings, also item intercepts are equal across groups. Scalar invariance implies that not only factor loadings, but also item intercepts are operating equivalently across groups. In models with ordered-categorical variables, instead of intercepts, thresholds are restricted testing for threshold invariance.

If this testing procedure reveals noninvariant items either on a metric or a scalar/threshold level, it cannot be assumed that the items of the translated instrument measure the same factor equally across groups. Occurrence of noninvariant item parameters relate to biases that can become evident on any level of equivalence and stem from differences in the measured constructs (construct bias), methodological artifacts such as familiarity with the item format (measurement bias) or problematic item content leading to cross-cultural differences in the interpretation of the item (item bias) (Borsboom, 2006; Byrne, 2016; van de Vijver & Tanzer, 2004).
Byrne, Shavelson, and Muthén (1989) introduced the approach of partial measurement invariance as an approach to treat noninvariant parameters. Here, some parameters are still constrained equal across groups whereas noninvariant parameters are identified and estimated freely (Byrne, 2012; Chen, 2008; Millsap & Kwok, 2004; Sass, 2011; Steinmetz, 2013). Alternatively, assembly can be considered as a strategy to deal with noninvariance.

Even though the issue of measurement invariance has been discussed in cross-cultural research for some decades (van de Vijver & Leung, 1997; Vandenberg & Lance, 2000), measurement invariant instruments in cross-cultural research within clinical psychology are still an exception (Byrne, Stewart, Kennard, & Lee, 2007; Crockett, Randall, Shen, Russell, & Driscoll, 2005; Nuevo et al., 2009; Torres, Miller, & Moore, 2013). Examples are the Brief Symptom Inventory (Hoe & Brekke, 2009), the Patient Health Inventory (Ryan, Bailey, Fearon, & King, 2013), the General Behavior Questionnaire (Pendergast et al., 2015), or Beck’s Depression Inventory (Byrne et al., 2007; Canel-Çınarbaş, Çui, & Lauridsen, 2011; Dere et al., 2015; Nuevo et al., 2009; Whisman, Judd, Whiteford, & Gelhorn, 2013); Testing measurement invariance of the DASS-21 across four countries, the current research contributes to the cross-cultural applicability of this instrument and extends the small number of comparable cross-cultural assessment instruments for symptoms of the most widespread mental health problems.

**Depression, Anxiety, and Stress Scales**

Based on the tripartite model of anxiety and depression (Clark & Watson, 1991), Lovibond and Lovibond (1995a) developed the Depression, Anxiety, and Stress Scales (DASS) that consists of 42 items and three dimensions (DASS-42). The depression scale comprises symptoms of anhedonia and inactivity (e.g., “I couldn’t experience any positive feeling at all.”). The anxiety scale describes symptoms of physiological hyperarousal and specific anxiety (e.g., “I found myself in situations that made me so anxious I was most relieved when they ended.”). General distress as a common symptom of depression and anxiety disorders is conceptualized in the stress scale (e.g., “I found it difficult to relax.”). A 21-item short version of the DASS with comparable psychometric properties was proposed by Lovibond and Lovibond (1995a) and subsequently validated (DASS-21; Antony, Bieling, Cox, Enns, & Swinson, 1998; Crawford & Henry, 2003). The short version rates depressive, anxiety, and stress symptoms over the past week on three 7-item subscales using 4-point Likert scales from 0 (did not apply to me at all) to 3 (applied to me very much or most of the time).

The DASS have several practical advantages: First, they are brief and can be applied in clinical and nonclinical samples (Antony et al., 1998; Brown, Chorpita, Korotitsch, & Barlow, 1997; Gloster et al., 2008; Gomez, Summers, Summers, Wolf, & Summers, 2014; Page, Hooke, & Morrison, 2007). In clinical samples, the DASS-42 distinguishes between anxiety and mood disorder patients (Brown et al., 1997) and assesses symptoms of depression and anxiety comorbid to other disorders (Smith, Sullivan, Hopkins, & Douglas, 2004; Szabó, 2011; Taylor, Lovibond, Nicholas, Cayley, & Wilson, 2005; Wilson & Rapee, 2005). Second, the DASS-21 can be used as a routine clinical outcome measure (Ng et al., 2007). Besides, it can serve to classify whether a patient moved into the range of normal functioning (Ronk, Korman, Hooke, & Page, 2013). Finally, the DASS-21 and the DASS-42 can serve as screening instruments (Dahm, Wong, & Ponsford, 2013; Kok, de Haan, van der Meer, Najavits, & De Jong, 2015; Nieuwenhuijsen, de Boer, Verbeek, Blonk, & van Dijk, 2003). Bayram and Bilgel (2008) propose the DASS-42 as a screening instrument to implement adequate primary and secondary prevention measures. For English-speaking populations percentile norms from an Australian general adult population sample (Crawford, Cayley, Lovibond, Wilson, & Hartley, 2011) as well as for a large psychiatric outpatients sample (Davies, Caputi, Skarvelis, & Ronan, 2015) exist for both versions of the DASS. It allows researchers and practitioners to compare the individual score in reference to the general population for screening and monitoring purposes. It must be kept in mind that most of these studies were conducted in English-speaking countries and thus the percentile norms do not necessarily apply internationally. Despite this limitation, all of the practical implications delineated above are encouraging to promote the implementation of the DASS-21 across countries. As a pre-condition, the challenges of cross-cultural adaptation need to be considered; namely, the scales’ translation, within-country validation and testing for measurement invariance across countries.

The DASS have been translated into 42 languages (Lovibond, 2015), which makes the scales widely accessible for practitioners and researchers. Yet, studies examining psychometric properties for the translated questionnaires are rare. Implicitly, researchers seem to assume that psychometric properties of the original questionnaire are transferrable and comparable for the translated version (Borsboom, 2006). However, this might not always be the case. Psychometric properties need to be investigated within the new population to ensure that the translated version functions like the original instrument. A literature search on PubMed, ISI, Web of Science, and PsyCINFO in June 2016 revealed 14 published validations of the questionnaire excluding the original English version. Psychometric properties were reported for the DASS-42 in Brazilian Portuguese (Sárdã, Nicholas, Pimenta, & Asghari, 2008), Chinese (Chan et al., 2012), Dutch (Nieuwenhuijzen et al., 2003), Italian (Severino & Haynes, 2010), and Turkish (Hekimoglu et al., 2012; Uncu, Bayram, & Bilgel, 2007). Validated and published versions of the DASS-21 exist in Brazilian Portuguese (Vignola & Tucci, 2014), Chinese (Gong, Xie, Xu, & Luo, 2010; Wang et al., 2016), Farsi (Asghari, Saed, & Dibajnia, 2008; Bayani, 2010; Sahebi, Asghari, & Safari, 2005), Indian (Singh, Junnarkar, & Sharma, 2015), Italian (Bottesi et al., 2015), German (Nilges & Essau, 2015), Malaysian (Musa, Fadzil, & Zain, 2007; Ramli, Ramli, Kartini, & Rosnani, 2011; Ramli, Salmiah, & Nurul Ain, 2009), Nepali (Tonsing, 2014), Portuguese (Apóstolo, Mendes, & Azeredo, 2006; Vasconcelos-Raposo, Fernandes, & Teixeira, 2013), Serbian (Jovanovic, Gavrilov-Jerkovic, Zuljevic, & Brdaric, 2014), Spanish (Antúnez & Vinet, 2012; Bados, Solanas, & Andrés, 2005; Daza, Novy, Stanley, & Aversill, 2002), and Vietnamese (Tran, Tran, & Fisher, 2013). Almost all of the studies included values for internal reliability reported as Cronbach’s alpha. Principal component analysis or confirmatory factor analysis as well as correlations with other psychometric measures were investigated in the majority of the studies. Only a small number of studies reported comparisons between clinical and nonclinical samples.
The DASS-21 has been tested only twice across different national groups. A study that tested measurement invariance across Asian cultures, namely Indonesia, Malaysia, Singapore, Sri Lanka, Taiwan, and Thailand had to revise the stress scale as it did not seem appropriate to Asian cultures (Oei, Sawang, Goh, & Mukhtar, 2013). Even though the three-factor model fitted the data well, the DASS-21 was not invariant across the Asian countries. Results of the second study suggested strong measurement invariance for the DASS-21 in paired comparisons of Australian and Chinese, as well as Australian and Malaysian, and Australian and Chilean adolescent samples (Mellor et al., 2015). Therefore, it is evident that measurement invariance testing of the DASS-21 should be continued to enhance the validity and the applicability of the DASS-21 internationally. Up till now, measurement invariance has not been tested in Eastern European or Central Asian countries such as Poland and Russia. This study aims to close this research gap.

Aim of the Study

The present study seeks to test measurement invariance of the DASS-21 in Poland and Russia as two Eastern European countries in reference to the United Kingdom (U.K.) and the United States of America (U.S.) as two representative Western countries. Additionally, latent mean differences are compared for symptoms of depression, anxiety, and stress provided that the DASS-21 is scalar measurement invariant. The study adds to the existing evidence of the cross-cultural applicability of the DASS-21 and enhances the options for practical implementation and cross-cultural research in the respective countries.

Method

Procedure

The Ethical Committee of the Faculty of Psychology at the Ruhr-Universität Bochum in Germany formally approved the study. Population-based surveys were conducted in Poland, Russia, the U.K., and the U.S. The study aimed to assess a representative sample of the population in all of the respective countries. To accomplish representativeness, each sample was drawn from the residential population aged 18 years and above that was accessible via landline or mobile phones. Landline telephone numbers were stratified by providers. An independent social market and research institute conducted the survey. In Russia, the survey was conducted from December 2013 to February 2014 and resulted in a total number of 3,022 completed interviews. Two interviews had to be excluded because the denoted age did not match the year of birth stated. 3,007 interviews were conducted in Poland, the U.K., and the U.S. from June to October 2014. Kish selection grid was implemented to identify the member within a household to be interviewed. Prior to each interview, the interviewer introduced the study and gathered participants’ informed consent. Response rate estimates of the proportion of actual eligible cases out of cases of unknown eligibility by the American Association for Public Opinion Research standards (The American Association for Public Opinion Research, 2016) varied between 13.4% and 20.1%.

Data Analysis

Statistical analyses were conducted using SPSS version 21 (IBM Corporation, 2011), Mplus (Muthén & Muthén, 2015), and R (R Development Core Team, 2013). For the analysis in R, the statistical packages lavaan (Rosseel, 2012) and psych (Revelle, 2015) were applied.

Participant Characteristics

In Russia, the sample comprised 3,020 participants. The Polish, the U.K., and the U.S. samples approximately constituted 1,000 participants per country. This should be Table 1. shows the demographic characteristics of the samples in each country. Whereas variations in mean age and retirement status might represent differences of the populations, the disparity of the highest level of education is partly caused by slightly different methodological assessment of the educational level across countries.

Measures

Demographic variables. Demographic variables were gathered according to internationally recognized standards. Gender, age, marital status, and current employment status were assessed according to the recommendations of the Eurostat task force on social variables (Eurostat, 2007). Highest educational level was assessed differently across countries corresponding to the respective educational system (European Social Survey, 2012). Harmonization of educational levels was guided by the International Standard Classification of Education (ISCED; UNESCO Institute for Statistics, 2013). Because of the differing operationalization, the Russian educational levels did not perfectly match the ISCE-Classification and could be only categorized approximately.

Depression, Anxiety, and Stress Scales. The freely available English version of the DASS-21 was used in the United Kingdom and the United States (Antony et al., 1998; Henry & Crawford, 2005; Lovibond & Lovibond, 1995b). The Polish version of the DASS-21 was retrieved from the DASS website (Lovibond, 2015). For the purpose of another study, our team had developed a Russian version of the DASS-21 that was included in the Russian survey. The translation procedure followed principles of good practice proposed for patient-related outcome measures by Wild et al. (2005). A native Russian speaking person with fluency in German managed the translation process. The project manager as well as a Russian translator carried out forward translation. An independent native psychologist who was not involved in the forward translation process reconciled both versions in collaboration with the translators. Back translation was carried out by two independent German speakers who were second generation Russian immigrants. Back translations were reviewed and adjusted by an expert panel consisting of the project manager and a German as well as a Russian clinical psychology professor.
Delta parameterization was used because it is also recommended for ordered-categorical data (Muthén & Muthén, 2012).

### Single-group confirmatory factor analyses

As baseline model, we tested the hypothesized three-factor model of the DASS-21 in all countries separately (Crawford et al., 2011; Daza et al., 2002; Lovibond & Lovibond, 1995a; Musa, Fadzil, & Zain, 2007; Sinclair et al., 2012). For this reason, single-group confirmatory factor analysis (CFA) was conducted for each country separately (for more detail see: Byrne, 2008; Hirschfeld & Von Brachel, 2014). For purposes of statistical identification, the first observed variable of each latent scale was used as indicator variable and fixed to 1. The baseline model is depicted in Figure 1.

### Multigroup confirmatory factor analyses to test measurement invariance

For multigroup confirmatory factor analysis, we used the more commonly applied forward approach which means including sequentially additional model constraints (Bovaird & Koziol, 2012; Hirschfeld & Von Brachel, 2014; Sass, 2011). To include all possible model comparisons, we tested Poland as the first reference country in comparison to the U.K. and the U.S., as well as the U.K. as reference country compared with all other three countries, followed by Russia as reference country in comparison to the U.K. and the U.S., as well as the U.K. as reference country compared with the U.S. This procedure was implemented in all subsequent steps of the measurement invariance testing.

### Testing configural invariance

As depicted in the introduction, configural invariance was tested first. This means that the baseline models that were established in each country separately were combined to form the multigroup model and tested simultaneously (Byrne, 2008). The three factors and their loading patterns are thus the same across groups, but not yet constrained to be equal across the U.K. (15.3% missing values). All of the DASS-21 variables taken together, 1.5% of the values in Poland were missing, as were 9.2% in Russia, 3.5% in the U.K., and 5.1% in the U.S. All available information was used to estimate confirmatory factor analyses. The procedure was guided by Asparouhov and Muthén’s (2010) proposal for missing data with weighted least squares estimation which is based on pairwise deletion.

Descriptive statistics were calculated with R. Skewness and kurtosis were calculated for each sample separately. An absolute value larger than 2 for skewness or larger than 7 for kurtosis was considered as reference for substantial non-normality as is recommended for samples larger than 300 (Kim, 2013; West, Finch, & Curran, 1995). Besides, Mardia’s normalized estimate of multivariate kurtosis was calculated as a measure to test the normality assumption (Mardia, 1970). Cronbach’s alpha indicated internal reliability and was considered acceptable above $\alpha \geq 0.70$ (Cronbach, 1951).

### Testing measurement invariance

Measurement invariance of the DASS-21 across four countries was tested based on Jöreskog’s (1971) multigroup confirmatory factor analysis using Mplus (Muthén & Muthén, 2015). The approach was specified for ordered-categorical variables by Millsap and Yun-Tein (2004), Muthén and Asparouhov (2002), and Muthén and Christoffersson (1981). The ordinal nature of our data was taken into account following their recommendations: Consistent with Norton (2007) and Mellor et al. (2015), we estimated parameters using robust weighted-least-squares (WLSMV; Flora & Curran, 2004; B. O. Muthén, 1984). WLSMV is recommended to estimate thresholds if fewer than five response categories are given (Beauducel & Herzberg, 2006). Delta parameterization was used because it is also recommended for ordered-categorical data (Muthén & Muthén, 2012).

#### Table 1

Sociodemographic Variables of the Participants

<table>
<thead>
<tr>
<th>Variable</th>
<th>PL</th>
<th>RU</th>
<th>UK</th>
<th>US</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participants, n</td>
<td>1003</td>
<td>3020</td>
<td>1002</td>
<td>1002</td>
</tr>
<tr>
<td>Gender, % female</td>
<td>47.9</td>
<td>53.2</td>
<td>53.6</td>
<td>52.5</td>
</tr>
<tr>
<td>Age, $M (SD)$</td>
<td>44.6 (16.1)</td>
<td>43.2 (17.1)</td>
<td>58.2 (17.8)</td>
<td>51.2 (16.9)</td>
</tr>
<tr>
<td>Marital status$^c$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unmarried</td>
<td>24.4</td>
<td>27.1</td>
<td>22.2</td>
<td>21.8</td>
</tr>
<tr>
<td>Married/Legal partnership</td>
<td>61.9</td>
<td>53.2</td>
<td>46.8</td>
<td>53.8</td>
</tr>
<tr>
<td>Widowed</td>
<td>6.2</td>
<td>8.4</td>
<td>16.5</td>
<td>9.3</td>
</tr>
<tr>
<td>Divorced/Separated</td>
<td>7.4</td>
<td>9.7</td>
<td>14.3</td>
<td>13.7</td>
</tr>
<tr>
<td>Missing</td>
<td>.1</td>
<td>1.6</td>
<td>.3</td>
<td>1.5</td>
</tr>
<tr>
<td>Highest level of education$^c$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary education or below</td>
<td>18.8</td>
<td>.5</td>
<td>6.7</td>
<td>.5</td>
</tr>
<tr>
<td>Lower secondary education</td>
<td>27.5</td>
<td>3.2</td>
<td>8.9</td>
<td>2.9</td>
</tr>
<tr>
<td>Upper and post-secondary education</td>
<td>16.6</td>
<td>59.0</td>
<td>36.2</td>
<td>53.1</td>
</tr>
<tr>
<td>Tertiary education (e.g. BA$^a$, MA$^b$)</td>
<td>35.8</td>
<td>37.3</td>
<td>29.2</td>
<td>38.3</td>
</tr>
<tr>
<td>Doctoral degree</td>
<td>.9</td>
<td>—</td>
<td>3.8</td>
<td>2.4</td>
</tr>
<tr>
<td>Missing</td>
<td>.8</td>
<td>—</td>
<td>15.3</td>
<td>2.8</td>
</tr>
<tr>
<td>Current employment status$^c$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paid work</td>
<td>60.3</td>
<td>52.2</td>
<td>39.5</td>
<td>48.9</td>
</tr>
<tr>
<td>Education</td>
<td>7.0</td>
<td>6.7</td>
<td>2.1</td>
<td>3.3</td>
</tr>
<tr>
<td>Unemployed</td>
<td>6.4</td>
<td>7.9</td>
<td>3.3</td>
<td>4.4</td>
</tr>
<tr>
<td>Permanently sick or disabled</td>
<td>.7</td>
<td>—</td>
<td>3.5</td>
<td>7.3</td>
</tr>
<tr>
<td>Retired</td>
<td>20.8</td>
<td>23.9</td>
<td>44.8</td>
<td>23.0</td>
</tr>
<tr>
<td>Housework, looking after children, other</td>
<td>4.2</td>
<td>8.0</td>
<td>4.1</td>
<td>6.0</td>
</tr>
<tr>
<td>Other</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>4.4</td>
</tr>
<tr>
<td>Missing</td>
<td>.6</td>
<td>1.2</td>
<td>2.7</td>
<td>2.8</td>
</tr>
</tbody>
</table>

*Note.* PL = Poland; RU = Russia; UK = United Kingdom; US = United States of America.  
$^a$BA = Bachelor. $^b$MA = Master. $^c$Frequencies are presented in percent.
groups. To identify the multigroup model, factor means were fixed to 0 and scaling factors were fixed to 1 in all groups.

Testing metric invariance. Second, we tested metric invariance and constrained factor loadings to be equal across countries. More specifically, the factor loadings of the group that served as the reference group were estimated freely, whereas the factor loadings of the second, third, and fourth group were fixed to equal the factor loadings of the reference group (Byrne, 2008). The metric model was identified constraining the factor means of the reference group to 0 and constraining the scaling factors of the reference group to 1. The factor means and scaling factors of the other groups were estimated freely. Additionally, the first threshold of all variables as well as the second threshold of all indicator variables were constrained to be equal across groups (Millsap & Yun-Tein, 2004).

Testing threshold invariance. Finally, because of the ordinal structure of the present data threshold invariance was tested constraining all thresholds to be equal across countries. This model does not require additional restrictions for identification; only the factor mean was fixed to 0 and scaling factors were fixed to 1 in the reference group. If threshold invariance is met, scalar invariance can be assumed (Millsap & Yun-Tein, 2004).

Testing latent mean differences. Latent means can be compared after establishing threshold invariance (Byrne, 2012). The comparison was based on the model used to test threshold invariance. The latent means of one group were fixed to zero. This group served as the reference group for the comparison with the latent means of the other groups (Byrne, 2012). Hence, only latent mean differences can be interpreted. z scores and p values of the standardized model results are reported (L. K. Muthén & Muthén, 2012). Cohen’s d was calculated as the effect size measure (small effect: d ≥ .20, medium effect: d ≥ .50, large effect: d ≥ .80; Cohen, 1988).

Evaluating model fit. Typically, the Likelihood Ratio Test statistic that is expressed as chi-square ($\chi^2$) statistic is used to evaluate the discrepancy between the covariance matrix of the restricted model and the covariance matrix of the unrestricted model (Byrne, 2012). A major limitation of using $\chi^2$ is its sensitivity to large sample sizes which leads to oversized rejection rates (Byrne, 2012; Hirschfeld & Von Brachel, 2014). Therefore, we reported $\chi^2$ only for reasons of sufficiency. A combination of two goodness-of-fit indices was used to evaluate our model: Root Mean Squared Error of Approximation (RMSEA; Steiger & Lind, 1980) and Comparative Fit Index (CFI; Bentler, 1990). All indices were based on robust values corrected in accordance with the WLSMV estimator. A CFI value of 1 would indicate perfect model fit. On the contrary, the value of 0 indicates perfect model fit for RMSEA. To ensure good model fit across different estimation methods, both goodness-of-fit indices were taken into account for the model evaluation (Schmitt, 2011). We followed Hu and Bentler’s (1999) recommendations to evaluate model fit. According to them RMSEA < .06-.08 and CFI > .95 are cut-off criteria for a good model fit.

The model fit of a nested model was compared with a less restricted model with a scaled difference $\chi^2$ test statistic (Asparouhov & Muthén, 2006; Satorra & Bentler, 2001). But as described above $\chi^2$ is very sensitive to large sample sizes, therefore, $\chi^2$ difference values are almost always significant. This problem
prompted researchers (Chen, 2007; Cheung & Rensvold, 2002) to conduct Monte Carlo studies as a means to propose recommendations based on CFI and RMSEA difference values. ΔCFI and ΔRMSEA were obtained calculating the difference between the CFI values, or RMSEA values respectively, for the restricted model against the less restricted model. Therefore, we considered the drop in the CFI-value and the increase of the RMSEA-value to evaluate the change of model fit. A change of ΔCFI ≥ 0.10 accompanied by a change of ΔRMSEA ≥ 0.015 indicated a significant decrease in model fit and hence noninvariance (Chen, 2007).

Results

Descriptive Statistics

Boxplots presented as Supplemental Material 1 showed low symptom levels in all countries with the highest ratings on the stress scale. However, a number of outliers were found for all countries and for all scales. Table 2 shows the descriptive results by nation. According to Kim (2013) and West et al. (1995), skew and kurtosis indicated a normal distribution. However, Mardia’s multivariate estimate for multivariate skew, Depression Scale: Median = 2.20, Anxiety Scale: Median = 3.05, Stress Scale = .75, and kurtosis, Depression Scale: Median = 5.01, Anxiety Scale: Median = 6.07, Stress Scale = 3.23, indicated a significant deviation from normality for all DASS-21 items. Internal consistency was ≥ .80 for all subscales except for the anxiety subscale in Poland, α = .78. In all countries α was ≥ .90 for the total scale. All item correlations were significantly positive and ranged from r = .14, p ≤ .001, U.K. sample: Item 1 and Item 7, to r = .67, p ≤ .001, U.S. sample: Item 17 and Item 21. Supplemental Material 2 displays the detailed correlation tables.

Single-Group Confirmatory Factor Analysis

The results of the confirmatory factor analysis of the assumed three-factor model in each of the four countries suggested that the model was appropriate across all countries. Because the baseline model was well-fitting in all countries, model respecifications were not necessary. The model fit was shown by line by line for each country in the upper part of Table 3 (for factor loadings of each scale by country, see Table 4). All CFI values exceeded the proposed cut-off score of > .95 for a good model fit (Hu & Bentler, 1999). The model fit was also good according to RMSEA, being smaller than the suggested cut-off score in all samples (Hu & Bentler, 1999).

Multi-Group Confirmatory Factor Analysis

Polish versus Russian, U.K., and U.S. samples. Results of the measurement invariance testing sequences are depicted in the lower part of Table 3 for each group comparison separately. In the first multigroup model, the Russian, the U.K., and the U.S. samples were compared with the Polish sample that was set as the reference country. Within this comparison, the first line shows the test of configural invariance. All fit indices indicated an appropriate model fit. The DASS-21 was also metric invariant as CFI was above .95 and RMSEA was below .08 for the metric model. ΔCFI was < .010 and ΔRMSEA < .015, which indicated that the more restricted model was as good as the less restricted model (Chen, 2007). When testing for threshold invariance, fit indices were still good. ΔCFI and ΔRMSEA did not indicate a meaningful drop of the model fit, ΔCFI = .003, ΔRMSEA = .006 (Chen, 2007). Thus, the Polish DASS-21 was threshold measurement invariant in comparison to the Russian, U.K., and U.S. questionnaire.

Russian versus U.K. and U.S. samples. The second multigroup model compared the U.K. and the U.S. to Russia. Again, fit indices of the configural model indicated configural invariance. Also, the metric model fitted the data well according to the fit indices and taking into account ΔCFI and ΔRMSEA, the model fit was as good as the configural model. The third line presents the results for the threshold model indicating a good model fit. Neither ΔCFI nor ΔRMSEA exceeded .01 and thus the model fit can be considered equal to the metric model (Chen, 2007). In sum, threshold invariance can be assumed for the Russian version in comparison to the U.K. and the U.S. versions of the DASS-21.

U.K. versus U.S. samples. Finally, the U.K. sample was compared with the U.S. sample in the third multigroup model. The configural model fitted the data well and so did both the metric and the threshold models. ΔCFI and ΔRMSEA dropped no more than .01 neither between the configural and the metric nor between the metric and the threshold model, thus threshold invariance can be assumed in this comparison as well.

Latent Mean Comparisons

Polish versus Russian, U.K., and U.S. samples. Figure 2 shows the latent mean differences with the Polish sample’s latent mean fixed to zero serving as the reference country. Comparisons
of the latent mean differences of the depression scale indicated
significant differences between the Polish sample and U.K.,
\(z = -3.47, p = .001, d = .16\), and U.S. samples, \(z = -3.92, p < .001, d = .18\), but no difference was found between the Polish
and the Russian sample, \(z = -0.87, p = .387, d = .03\). The latent
mean of the anxiety scale was significantly higher in the Russian
sample than in the Polish sample, \(z = -3.47, p = .001, d = .13\),
whereas the differences between the Polish sample and U.K.,
\(z = -1.59, p = .113, d = .07\), and U.S. samples were not
significant, \(z = -0.50, p = .618, d = .02\). The Polish sample
scored significantly higher on the latent mean of the stress scale
compared with the Russian sample, \(z = -4.72, p < .001, d = .17\),
the U.K. sample, \(z = -4.49, p < .001, d = .20\), and the U.S.
sample, \(z = -3.99, p < .001, d = .18\).

#### Russian versus U.K. and U.S. samples.

In a second step, the Russian sample’s latent mean was fixed to zero.
Thus, the latent mean of the Russian sample could be compared with the U.K.
and the U.S. sample. On the depression scale, the Russian sample
scored significantly higher than the U.K. sample, \(z = -3.70, p < .001, d = .14\), and the U.S. sample, \(z = -4.15, p < .001, d = .15\).
The latent mean of the anxiety scale was significantly higher in the
Russian sample than in U.K., \(z = -6.10, p < .001, d = .23\), and
U.S. samples, \(z = -4.26, p < .001, d = .16\). Finally, the com-
parisons between stress symptoms in the Russian sample and the
U.K. sample, \(z = -0.91, p = .363, d = .03\), and the U.S. sample,
\(z = -0.44, p = .65, d = .02\), did not yield significant differences.

#### U.K. versus U.S. samples.

The latent mean of the U.K. sample
was fixed to zero to conduct the latent mean comparison
between the U.K. and the U.S. sample. U.K. and U.S. samples did
not differ from each other significantly neither on the depression,
\(z = -0.72, p = .472, d = .03\), and anxiety scales, \(z = -0.52, p = .604, d = .02\), nor on the stress scale, \(z = 0.20, p = .841, d = .01\).

#### Discussion

The present study examined the measurement invariance of the
DASS-21 across randomly selected samples in Poland, Russia, the
U.K., and the U.S. Threshold measurement invariance could be
established across all countries. Thus statistical requirements for

---

### Table 3

**Fit-Indices for Single- and Multi-Group Confirmatory Factor Analyses**

<table>
<thead>
<tr>
<th></th>
<th>CFA (\chi^2(df)^a)</th>
<th>CFI</th>
<th>RMSEA (90% CI)</th>
<th>(\Delta \chi^2) b</th>
<th>(\Delta CFI)</th>
<th>(\Delta RMSEA)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Single-group CFA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PL</td>
<td>854.520 (186)</td>
<td>.962</td>
<td>.060 (.056; .064)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RU</td>
<td>2323.379 (186)</td>
<td>.964</td>
<td>.062 (.059; .064)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UK</td>
<td>814.666 (186)</td>
<td>.967</td>
<td>.058 (.054; .062)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>US</td>
<td>715.747 (186)</td>
<td>.982</td>
<td>.053 (.049; .057)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Multi-group CFA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1) PL vs. RU vs. US</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Configural invariance</td>
<td>4548.319 (744)</td>
<td>.970</td>
<td>.058 (.057; .060)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metric invariance</td>
<td>5770.799 (798)</td>
<td>.961</td>
<td>.064 (.063; .066)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Threshold invariance</td>
<td>5564.147 (915)</td>
<td>.964</td>
<td>.058 (.057; .060)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2) RU vs. UK vs. US</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Configural invariance</td>
<td>3729.319 (558)</td>
<td>.971</td>
<td>.058 (.056; .060)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metric invariance</td>
<td>4847.355 (594)</td>
<td>.961</td>
<td>.065 (.064; .067)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Threshold invariance</td>
<td>4774.726 (672)</td>
<td>.962</td>
<td>.060 (.059; .062)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(3) UK vs. US</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Configural invariance</td>
<td>1533.404 (372)</td>
<td>.976</td>
<td>.056 (.053; .059)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metric invariance</td>
<td>1578.049 (390)</td>
<td>.976</td>
<td>.055 (.052; .058)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Threshold invariance</td>
<td>1453.904 (429)</td>
<td>.979</td>
<td>.049 (.046; .052)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* CFA = confirmatory factor analysis; PL = Poland; RU = Russia; UK = United Kingdom; US = United States of America.

With one exception (see below), all \(\chi^2\) values are significant, \(p \leq .000\). b The test was based on robust \(\chi^2\) difference testing (calculation based on Asparouhov & Muthen, 2006). *\(\chi^2\) value is not significant, \(p = .11\).*

### Table 4

**Standardized Factor Loadings of Each Subscale of the DASS-21 Within Each Country**

<table>
<thead>
<tr>
<th>Scale</th>
<th>PL</th>
<th>RU</th>
<th>UK</th>
<th>US</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depression</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dass3</td>
<td>.66</td>
<td>.69</td>
<td>.72</td>
<td>.72</td>
</tr>
<tr>
<td>dass5</td>
<td>.59</td>
<td>.68</td>
<td>.68</td>
<td>.74</td>
</tr>
<tr>
<td>dass10</td>
<td>.76</td>
<td>.57</td>
<td>.86</td>
<td>.89</td>
</tr>
<tr>
<td>dass13</td>
<td>.82</td>
<td>.86</td>
<td>.86</td>
<td>.90</td>
</tr>
<tr>
<td>dass16</td>
<td>.83</td>
<td>.76</td>
<td>.85</td>
<td>.86</td>
</tr>
<tr>
<td>dass17</td>
<td>.82</td>
<td>.77</td>
<td>.88</td>
<td>.88</td>
</tr>
<tr>
<td>dass21</td>
<td>.85</td>
<td>.75</td>
<td>.91</td>
<td>.94</td>
</tr>
<tr>
<td>Anxiety</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dass2</td>
<td>.50</td>
<td>.55</td>
<td>.61</td>
<td>.62</td>
</tr>
<tr>
<td>dass4</td>
<td>.64</td>
<td>.65</td>
<td>.62</td>
<td>.72</td>
</tr>
<tr>
<td>dass7</td>
<td>.68</td>
<td>.73</td>
<td>.64</td>
<td>.72</td>
</tr>
<tr>
<td>dass9</td>
<td>.69</td>
<td>.77</td>
<td>.79</td>
<td>.85</td>
</tr>
<tr>
<td>dass15</td>
<td>.90</td>
<td>.88</td>
<td>.93</td>
<td>.90</td>
</tr>
<tr>
<td>dass19</td>
<td>.72</td>
<td>.69</td>
<td>.69</td>
<td>.78</td>
</tr>
<tr>
<td>dass20</td>
<td>.82</td>
<td>.84</td>
<td>.86</td>
<td>.90</td>
</tr>
<tr>
<td>Stress</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dass1</td>
<td>.59</td>
<td>.64</td>
<td>.67</td>
<td>.69</td>
</tr>
<tr>
<td>dass6</td>
<td>.66</td>
<td>.71</td>
<td>.72</td>
<td>.73</td>
</tr>
<tr>
<td>dass8</td>
<td>.74</td>
<td>.82</td>
<td>.81</td>
<td>.88</td>
</tr>
<tr>
<td>dass11</td>
<td>.79</td>
<td>.79</td>
<td>.83</td>
<td>.85</td>
</tr>
<tr>
<td>dass12</td>
<td>.79</td>
<td>.81</td>
<td>.84</td>
<td>.80</td>
</tr>
<tr>
<td>dass14</td>
<td>.71</td>
<td>.76</td>
<td>.64</td>
<td>.74</td>
</tr>
<tr>
<td>dass18</td>
<td>.74</td>
<td>.80</td>
<td>.80</td>
<td>.85</td>
</tr>
</tbody>
</table>

*Note.* PL = Poland; RU = Russia; UK = United Kingdom; US = United States of America.
comparisons of the relationships between latent variables across groups as well as latent mean comparisons are fulfilled.

Scale means were rather low in all samples indicating low symptom levels in the general population. However, the boxplots of the samples demonstrated extreme responses indicating higher symptom levels in a subset of all samples. These subsets might represent the part of the population suffering from more severe mental distress, which would be in line with the high prevalence rates outlined in the introduction. Generally, our findings correspond to the results of Henry and Crawford (2005) and are slightly lower than results from other representative samples (Crawford & Henry, 2003; Lovibond & Lovibond, 1995b; Ronk et al., 2013). This difference could be attributable to the assessment methods. Telephone interviews probably cause more socially desirable behavior and thus result in lower scale scores (Zhang, Kuchinke, & Margraf, 2015).

The results testing the normality assumption were mixed. According to the cut-off scores proposed by Kim (2013) and West et al. (1995), the data were normally distributed, whereas based on Mardia’s normalized estimate of multivariate kurtosis the data indicated non-normality in all countries. As the data were left skewed, a moderate non-normality is assumed. Because all analysis were based on WLSMV estimator, such moderate non-normality should not affect the accuracy of the estimation (Flora & Curran, 2004).

Concerning the reliability of the DASS-21, internal consistency was excellent for the total scale and good for the depression, anxiety, and stress scales, respectively. Merely, Cronbach’s alpha of the anxiety scale was only acceptable in the Polish sample. These findings, especially the lower internal consistency of the anxiety scale, correspond to the results of previous studies (Antúnez & Vinet, 2012; Bados, Solanas, & Andrés, 2005; Crawford et al., 2011; Daza et al., 2002; Musa et al., 2007; Vasconcelos-Raposo, Fernandes, & Teixeira, 2013).

Within each country, the appropriateness of the theoretically assumed three-factor model was tested. Single-group confirmatory factor analyses showed that the three-factor model fitted the data in all countries. Multigroup confirmatory factor analyses showed that the assumptions of equivalent loadings and thresholds held across all countries. However, it is puzzling that model fit did not decrease from configural to metric and scalar models, as is common in measurement invariance testing using Maximum Likelihood estimation. Instead, χ² indicated a better model fit for the more restricted model in all country comparisons. This pattern was also found in the multigroup comparison of the DASS-21 across four racial groups conducted by Norton (2007) who also used WLSMV. Also, ΔCFI and ΔRMSEA did not decrease, nor increase, respectively, as one would expect for the more restricted models. For example, with Poland as the reference country, the CFI was higher in the configural model than in the metric model as expected, but it was also higher in the threshold model compared with the metric model. It is most probable that this pattern is caused by the model identification that is needed in the case of ordered-categorical variables. On each level of measurement invariance testing, different model parameters have to be restricted to identify the model (Millsap & Yun-Tein, 2004; see also the paragraph “Multi-group confirmatory factor analyses to test measurement invariance” in the Method section). Other studies using WLSMV have reported similar patterns (Bieda et al., 2016; Mellor et al., 2015; Norton, 2007), even though the results were not discussed in detail. Further research on multi-Group CFA with ordered-categorical variables should be conducted to better understand the behavior of the Likelihood Ratio Test statistic as well as of the goodness-of-fit indices with ordered-categorical variables. In the present results, all threshold models indicated excellent model fit that did not deviate significantly from the metric models. Hence, these results suggest that relationships between latent variables and latent means can be meaningfully compared across countries.

The results of the latent mean comparisons indicated a clear difference between the two Eastern European and the two Western countries. Whereas the U.S. and the U.K. did not differ in regard to symptoms of depression, anxiety, and stress, both Polish and Russian samples showed more symptoms of depression. However, all effect sizes were very small. High symptom levels of depression are in line with higher prevalence rates of depression found in the Czech Republic, Russia, and Poland (Bobak et al., 2006). One reason for the difference could be the less prosperous economies in Russia and Poland compared with the U.S. and the U.K. Rai, Zitko, Jones, Lynch, and Araya (2013) found that symptoms of depression vary up to 13.5% at the country-level and that symp-
toms increase with decreasing levels of economic development. More symptoms of anxiety were only reported for the Russian sample compared with the other three countries which did not differ significantly in their anxiety levels. In terms of effect sizes, the Polish and U.S. samples did not differ from the Russian sample, but a small effect was present for the difference between the U.K. and the Russian sample. This finding is in line with very high levels of anxiety found in another Russian sample (Averina et al., 2005) and our findings yield to the fact that this might not be a measurement error. Also prevalence rates of social anxiety have been found to be higher in Russia (Eaton et al., 2008). Such relatively high levels of anxiety might be attributable to rising inequalities in Russia caused by the new globalized neoliberal economic conditions (Burns, 2015). Surprisingly, not the Russian, but the Polish sample reported the highest levels of stress compared with the Russian, U.K., and U.S. sample. A small effect was found compared with the U.K. sample, but effect sizes in comparisons with the Russian and the U.S. samples were very small. Watson (2006) argues that high levels of stress in Poland are linked to a lack of security of employment, lower income, increasing inequality and changes in social relations. It might be particularly burdensome to the Polish population that the country quickly realized its transition to become a democracy and member of the European Union (Carlson, 2016). Now, Poland competes with economically strong countries such as Germany, France, or the U.K., which might increase the pressure and stress levels in people. All of the outlined explanations for latent mean differences across countries should be viewed as a part of a larger picture. Future studies are needed to replicate our findings and test our hypothesis more explicitly to validate or falsify our conclusions.

Generally, such comparisons and hypotheses about cultural differences require prior testing of measurement invariance. The subsequent considerations are very valuable for the comparison of symptoms of mental distress across countries and for the understanding of the social and cultural circumstances behind these differences. This underlines the importance of measurement invariance testing.

Methodological Challenges and General Limitations

Numerous articles and reviews have been published on measurement invariance (e.g., Byrne, 2008; Cheung & Rensvold, 1999; Milfont & Fischer, 2010; Steenkamp & Baumgartner, 1998; van de Schoot, Lugtig, & Hox, 2012; Vandenberg & Lance, 2000). Despite such a large body of research, recommendations for multigroup confirmatory factor analysis with ordered-categorical variables are rare and do not seem to be as straightforward as recommendations for metric variables (Bovaird & Koziol, 2012; Hirschfeld & Von Brachel, 2014; Millsap & Yun-Tein, 2004; B. O. Muthén & Asparouhov, 2002; B. O. Muthén & Christoffersson, 1981). For example, whereas the minimal model identification constraints proposed by Millsap and Yun-Tein (2004) are considered the gold-standard for model identification with ordered-categorical variables, the authors themselves admit that the identification fails in some cases. Bovaird and Koziol (2012) propose a different approach including the restriction of residual variances and Hirschfeld and Von Brachel (2014) do not outline the model identification at all, instead they use the “group.equal” command within the lavaan package (Rosseel, 2012) for the group comparisons. For a clinically oriented researcher such differing approaches might be difficult to evaluate. Furthermore, only recently Sass, Schmitt, and Marsh (2014) gave an overview of the evaluation of model fit with ordered categorical data. Concerning ΔCFI and ΔRMSEA that indicate a meaningful drop of model fit, simulation studies have used maximum-likelihood-estimation (Chen, 2007; Cheung & Rensvold, 2002). It remains unknown whether the cut-off is also a good indicator for models estimated with WLSMV (Hirschfeld & Von Brachel, 2014). Future simulation studies should consider ordered-categorical variables in respect to model identification, estimation, and evaluation in further detail. And most importantly, results should be summarized in comprehensive reviews that can guide future analysis.

Concerning general limitations, only the internal structure of the DASS-21 was analyzed and compared across countries. This leaves other aspects of reliability and validity unanswered and limits the scope of the present study. Nevertheless, as outlined in the introduction, numerous studies have examined the reliability and validity of the DASS-21.

The focus on general population samples across four countries brings about a second general limitation: A part of the population might not be accessible via landline or mobile phones (Brick, Dipko, Presser, Tucker, & Yuan, 2006). Additionally, the response rates indicate that only a small proportion of the generally accessible population agrees to participate in a survey. To deal with accessibility bias and nonresponse, the cases can be weighted according to the respective general population to generalize results (Little & Rubin, 2002). Even though the possibility to weigh cases was available for the present study as well, it was not used in the analyses because weighing cases would mean that the responses of some cases are multiplied and other responses are calculated less than once to fit the gender, age, and regional structure of the population under study. For instance, it is assumed that statistically inflated cases represent the respective population, for example, men aged 34–40 years living in Berlin. However, this assumption could lead to false conclusions based on data with weighted cases (Halbesleben & Whitman, 2013). Nevertheless, only very few studies on the DASS-21 report large samples with the attempt of representativeness (Crawford et al., 2011; Henry & Crawford, 2005; Lovibond & Lovibond, 1995a; Ronk et al., 2013). Thus, despite the limitation of not demonstrating perfect representativeness, from the authors’ perspective, we consider the sample large and broad enough to generalize our findings.

A third limitation concerns the method of data collection. The DASS-21 was developed as a paper-pencil questionnaire. In the present study, item responses were assessed verbally in a telephone interview. Altering the survey method might affect social desirability bias (Persoskie, Leyva, & Ferrer, 2014). Another study conducted by our research team (Zhang et al., 2015) compared differences of online, paper-pencil, face-to-face and telephone interviews and the results showed lower severity reports in telephone interviews compared with online or paper-pencil. This would support the hypothesis of the influence of social desirability. However, the effects were smaller than η² = 0.04 and the assessment with the DASS-21 was strongly measurement invariant across all survey methods. Still, it must be kept in mind that responses might be generally lower than they would be in paper-pencil format.
Fourth, measurement invariance is a statistical procedure testing the comparability of an instrument across groups. Yet, it is difficult to disentangle actual cultural differences from the variety of possible biases and artificial errors in the measurement, such as social norms or data collection methods (Byrne, 2008). Therefore, testing for measurement invariance is only one step toward culturally sensitive measurement instruments. Qualitative research approaches such as cognitive interviewing (Beatty & Willis, 2007) could be used in the translation process to address possible construct and item bias, but also in post hoc analysis to explore reasons for noninvariance.

**Conclusion**

Facing the burden of an increasing number of mental disorders, researchers worldwide need to work together to develop, exchange, and evaluate prevention and treatment strategies. For collaboration in research and practice, reliable and valid assessment instruments that enable the measurement and comparison of psychological symptoms are fundamental. Cross-cultural research and practice should address additional methodological questions concerning measurement invariance of the assessment instruments as well as questions regarding real cultural differences. This study presents tests of measurement invariance of the DASS-21 that was shown to be threshold measurement invariant in Poland, Russia, the U.K., and the U.S. The analyses of relationships between latent variables as well as latent mean comparisons are allowed based on the full scales across all investigated countries. Future studies in cross-cultural settings need to test measurement invariance in other clinical assessment instruments to overcome the challenge of assessing psychological symptoms across nations and to avoid bias due to the interpretation of distorted observed scale means. More methodological research on measurement invariance testing with ordered-categorical variables is needed to clarify and standardize the procedures of the analysis.

**References**


Bowden, A., & Fox-Rushby, J. A. (2003). A systematic and critical review of the process of translation and adaptation of generic health-related quality of life measures in Africa, Asia, Eastern Europe, the Middle East,.
South America. Social Science & Medicine, 57, 1289–1306. http://dx.doi.org/10.1016/S0277-9536(02)00503-8


MEASUREMENT INVARIANCE OF THE DASS-21


Received December 4, 2015
Revision received December 2, 2016
Accepted December 7, 2016