Dimensionality and Measurement Invariance of the Serbian Version of the BDI-II:

An IRT Approach

*Corresponding author: janicic@ff.uns.ac.rs

This is an early electronic version of the manuscript that has been accepted for publication in Psihologija journal but has not yet been technically prepared for publication. Please note that this is not the final version of the paper as it has yet to be technically prepared for publication and minor changes to the text are possible before the final print. The final version of the article can be subjected to minor changes after proof reading and before final print. Please cite as: Janičić, B., Mihić, Lj., & Novović, Z. (2020). Dimensionality and Measurement Invariance of a Serbian Version of the BDI-II: An IRT Approach. Psihologija. Advance online publication. doi: https://doi.org/10.2298/PSI200318027J
There have been debates about the dimensionality of the Beck’s Depression Inventory – II, its appropriate scoring, and gender-related measurement invariance. We addressed these questions employing the Item Response Theory approach in a clinical sample of 288 patients, using a Serbian version of the BDI–II. We tested nine structural models using confirmatory Full Information Factor Analysis and IRT Graded Response Model. We concluded that the BDI-II is essentially unidimensional. All items had high discrimination, and the test was most informative in the medium range of depression severity in the clinical sample. Although small to medium gender related differential item functioning existed in several items, it did not affect the total score. Hence, the total score of the Serbian translation of the BDI–II is comparable between genders as a measure of severity of depression.

**Key words:** depression, BDI-II, Item Response Theory, Differential Item Functioning, gender differences

**Highlights:**

- The Serbian version of the BDI-II has essentially unidimensional structure.
- The BDII–II is highly informative for measuring depression in clinical samples.
- Small to moderate DIF exists in items measuring vegetative symptoms and crying.
- Severity of depression, as the total BDI–II score, is comparable between genders.
One of the most popular and widely used self-assessment questionnaires of depression is the Beck Depression Inventory–II (BDI–II; Beck et al., 1996; Mihić & Novović, 2019). Both versions, BDI–IA (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961) and the actual, second edition, have been thoroughly psychometrically analyzed, and good reliability and validity have been consistently reported (Erford, Johnson, & Bardoshi, 2016).

Factor structure of the BDI–II has been reported in various psychometric studies, but the results have not been uniform. Depending on sample type and size, type of factorization and methodology, different combinations of the affective, cognitive, and somatic factors have been identified i.e., Cognitive and Somatic-Affective (Beck et al., 1996), Cognitive-Affective and Somatic-vegetative (Dozois, Dobson, & Ahnberg, 1998; Patterson et al., 2011; Seignourel, Green, & Schmitz, 2008), Cognitive-Somatic and Affective (VanVoorhis & Blumentritt, 2007), and Cognitive-Affective and Exhaustion (Novović, Mihić, Tovilović, Jovanović, & Biro, 2011). The three- and even the four-factor solutions have also been reported (Beck, Steer, Brown, & Van der Does, 2002; Buckley, Parker, & Heggie, 2001; Rodriguez-Gomez, Dávila-Martínez, & Collazo-Rodríguez, 2006; Subramaniam, Harrell, Huntley, & Tracy, 2009). As some authors noted, in most studies the affective symptoms appeared unstable, switching between the cognitive and the somatic factors in different studies (e.g., Bühler, Keller, & Läge, 2014).

It is interesting that one-factor solution has been rarely supported (Kim, Pilkonis, Frank, Thase, & Reynolds, 2002; Segal, Coolidge, Cahill, & O’Riley, 2008). However, recently, several bifactor analytic studies have re-vitalized an interest in the general factor (G). These studies have agreed that the BDI–II has a bifactor structure with a dominant influence of the G factor on the common BDI–II item variance in comparison to specific factors. Even though these studies differed in the number and type of specific factors extracted – some reported two (Al-Turkait & Ohaeri, 2010; de Miranda Azevedo et al., 2016; Osman, Barrios, Gutierrez, Williams, & Bailey,
2008; Subica et al., 2014; Ward, 2006) and some three factors (Bühler et al., 2014; Quilty, Zhang, & Bagby, 2010) – they all have agreed that “...G is the key” (Brouwer, Meijer, & Zevalkink, 2013, p. 136). Hence, a majority of authors would argue that multidimensionality within the BDI–II items stems from different item content domains rather than the presence of specific factors that deserve to be scored separately (but see Bühler et al., 2014 for a different interpretation of the importance of the specific factors). Therefore, the instability of the factor solutions might stem from an incorrect model specification, in particular, a strong presence of the general factor (G) in all BDI–II items that could not be specified in studies that did not apply bifactor modeling (Ward, 2006). Namely, some items that are strongly saturated by the G factor might switch places from one oblique factor to the other in different studies (Ward, 2006). Others suggested sample-dependent error variance as an explanation for various factor analytic solutions (Strunk & Lane, 2017).

One goal of the present study is to provide a comprehensive test of the BDI–II structure in a sample of depressed patients from Serbia. Based on the reviewed studies, we expected that most of the common BDI–II item variance would be explained by the dominant G factor. If supported, this finding would inform clinicians about the best way of scoring the instrument when used for screening and research purposes. In particular, the Serbian version of the BDI would be scored as the total score, which is the BDI-II original scoring.

It is also possible that not all BDI–II items function uniformly (e.g., discriminate equally) across the latent trait of severity of depression and/or across different groups of individuals. For example, item difficulty (i.e., locations of categories’ thresholds in polytomous items) might differ among various demographic groups, which might suggest a presence of differential item functioning (DIF) or a lack of measurement invariance. DIF exists when individuals, belonging to different demographic groups but with the same latent trait level (theta), respond differently to
a test item. Such item characteristics can be tested within the item response theory framework (IRT), which has already been employed several times in testing psychometric characteristics of the BDI–II (de Sá Junior, de Andrade, Andrade, Gorenstein, & Wang, 2018; de Sá Junior et al., 2019; Kim et al., 2002; Wu & Huang, 2010). If our structural analyses show that the BDI-II is highly saturated by the G factor, we will fit unidimensional IRT model with two aims: to better understand how the BDI-II items function as a measure of severity of depression, and whether there is differential item functioning (DIF) between genders.

In IRT models, specifically the graded response model – GRM (Samejima, 1969), the parameter $a$ (slope) together with the categories' location parameters ($b$), indicate how well item response categories (0–3 in the case of the BDI–II) discriminate between the levels of severity of depression considered as a latent trait. In GRM, the categories’ location parameters (thresholds) indicate the point on the latent trait where the probability of endorsing neighboring response categories is equal (.50) (Reise et al., 2011). The thresholds can help one determine whether the item or test difficulty is adequate for the primary target population. For example, in a large sample of Brazilian university students, it has been shown that all BDI–II items have moderate-to-high discrimination parameters for the latent trait of severity of depression (de Sá Junior et al., 2019). To our knowledge, there is no study that reported detailed $a$ and $b$ parameters for the BDI-II items in a sample of clinically depressed individuals.

Only two studies explored DIF between genders. In a student sample, de Sá Junior et al. (2019) concluded that BDI–II items #10 (Crying) and #21 (Loss of interest in sex) demonstrated DIF, with women having a greater tendency to endorse these items, however, the overall scores were negligibly impacted by such differences. In a clinical sample and using the BDI–IA, Santor Ramsay, and Zuroff (1994) reported that only #14 (Distortion of body image) tends to
overestimate the degree of depression in women compared to men. Thus, there is clearly a need for a replication study on an independent patient sample.

The importance of disentangling item bias from real differences between genders is particularly great, given inconsistent findings regarding gender differences registered via the BDI–II or its earlier version. For example, a number of studies reported that women obtained higher scores than men in different populations (e.g., in adult clinical population, see Beck, Steer, & Brown, 2009). However, differences were not always reported (e.g., in adult general population see Richter, Polak, & Eisemann, 2003).

Hence, the aims of the present paper were: to provide a comprehensive assessment of various factor analytic models of the BDI–II (Serbian version) that would inform us about optimal scoring of BDI-II, and, contingent upon its hypothesized unidimensionality, to examine item-level IRT characteristics on a sample of Serbian depressed patients and gender-related DIF.

Method

Sample

The sample consisted of 288 psychiatric patients (36.8% male) recruited from 9 psychiatric hospitals in different cities across Serbia, with an average age of $M = 49.90$ ($SD = 11.20$). There was no significant age difference ($t_{(269)} = .78$, $p = .433$, Cohen’s $d = .09$) between males and females ($M_m = 50.64$, $SD_m = 11.96$, $M_f = 49.49$, $SD_f = 11.20$).

Patients were classified according to the ICD–10 (World Health Organization, 1992) by one mental health professional based on clinical interviews, clinical psychological testing, and case notes. Patients had the following primary diagnoses: depressive episode (37%), recurrent depressive disorder (38%), dysthymia (3%), unspecified affective disorder (4%), brief depressive adjustment reaction (1%), prolonged depressive adjustment reaction (1%), mixed anxiety and
depressive reaction (5%), and bipolar affective disorder - current depressive episode (12%). 46% patients had co-morbid diagnoses. Psychosis and dementias were exclusionary criterion.

**Instruments**

The Beck Depression Inventory–II (BDI–II; Beck et al., 1996; Mihić & Novović, 2019) is a multiple-choice, 21-item self-report measure of severity of depression. A previous study supported validity and reliability of the translated version of the BDI–II (Mihić & Novović, 2019). Each answer is scored on a scale ranging from 0 to 3. The total score is the sum of all item scores. In the present study, a coefficient of reliability was high, $\alpha = .94$.

**Data Analysis**

Data were analyzed in the R environment (R Core Team, 2019) using packages *mvoutlier* (Filzmoser & Gschwandtner, 2018), *psych* (Revelle, 2009), *ltm* (Rizopoulos, 2006) and *mirt* (Chalmers, 2012). Taking into account a large number of different factor solutions reported in the literature to date and the ordinal nature of the data, it would be useful to utilize IRT full information confirmatory factor analysis (FA) approach (e.g., Bock, Gibbons, & Muraki, 1988). This method of FA uses frequencies of distinct response vectors as data (Bock et al., 1988). All fitted models were two-parameter logistic graded response models (Samejima, 1983). These models estimate location ($b$ – difficulty) parameters for every threshold between items’ categories and one slope ($a$ – discrimination) parameter per item. In addition, IRT analysis produces information functions of items and test, and can be used to examine measurement invariance (Meade, 2016) through analyses of differential item and test functioning (DIF and DTF).

**Results**
Descriptive Characteristics

Average BDI–II total score was $M = 26.74$ ($SD = 14.49$). Skewness of the total score was negligible ($Sk = .105$, $SE_{Sk} = .144$), and kurtosis was negative and moderate ($Ku = -.856$, $SE_{Ku} = .286$). Descriptive statistics for each item are presented in Table 1. There was no significant gender difference in the BDI–II total score ($t_{(286)} = -1.89$, $p = .060$, Cohen’s $d = -.23$). Mean score for females was $M = 27.97$ ($SD = 14.34$), and for males $M = 24.44$ ($SD = 14.46$).

Latent Structure of the BDI–II

In order to determine the latent structure of the Serbian version of the BDI–II, nine structural models were tested. To provide continuity with the previous studies, we explored the models tested in Brouwer et al.’s study (2013), which at the time represented the most comprehensive empirical test of various BDI–II factor structures. We also included the bifactor model that was published subsequently (Bühler et al., 2014) and a two-factor correlated model previously reported on a Serbian, university sample (Novović et al., 2011). Detailed specifications of all models are presented in Table 2.

We considered $CFI \geq .95$ (acceptable if $.90 \leq CFI < .95$), $TLI \geq .95$, $RMSEA \leq .05$ (acceptable if $.06 < RMSEA < .08$), and $SRMSR \leq .09$ (Hu & Bentler, 1999) as indicators of good model fit. Since more than half of tested models were bifactor, and global fit indices tend to favor bi-factor models due to their tendency to overfit (Bornovalova, Choate, Fatimah, Petersen, & Wiernik, 2020; Markon, 2019) we carefully inspected all the solutions, including loadings on the general and group factors. Apart from the model fit, we considered additional criteria while
evaluating models: a) proportion of uncontaminated correlations (PUC) i.e., the proportion of all possible item correlations which are not contaminated by the correlations among items that belong to the same group factors, b) proportion of explained common variance (ECV), c) factor determinancy (FD), and d) replicability index H. FD (correlation between factor and factor scores) is proposed to be satisfactory if >.90 (Gorsuch, 1983). H index is a measure of factor replicability, with the values > .80 suggesting well-defined latent variables (Hancock & Mueller, 2001). Finally, $\omega_h$ coefficient estimates how much of the total BDI–II score variance is attributable to the G factor, whereas $\omega_{hs}$ reflects systematic variance that is left once individual variability due to the general factor was partitioned (Reise, Scheines, Widaman, & Haviland, 2013). The value of $\omega_h > .80$ suggests that the scale can be regarded as measuring a unidimensional construct (Reise et al., 2013).

Limited information goodness of fit statistic $\chi^2$ (Table 3) was significant in all but three models (M5, M9, and M6) indicating lack of fit. $\chi^2$ statistic tests for exact fit and it tends to be significant in models with large degrees of freedom (Maydeu-Olivares, 2013). Hence, we assessed approximate model fit based on other indices. More than a half of the tested models satisfied all the criteria for good fit (Table 3).

INSERT TABLE 3 HERE

According to Akaike Information Criterion (corrected for sample size), the best fitting model is M3 (the three-correlated factors model), followed by models 5, 2, 4, 9 (Table 3). The differences between the models’ fit were also tested using likelihood ratio tests \(^2\). The best fitting models according to LR tests were 5 and 9 i.e., the two bifactor models which differed mainly in their definition of one group factor (Cognitive, Somatic, and Affective vs. Cognitive, Somatic, and Activation). However, inspection of standardized item loadings in both models (Table 4)

\(^2\) Details available upon request from the first author.
revealed an inadequate representation of group factors judged by the number of significant loadings and presence of small negative loadings, indicating model misspecification (Reise, Kim, Mansolf, & Widaman, 2016). Also, all other indices in Table 4 were adequate for the general factor only, supporting rejection of the bifactor models.

INSERT TABLE 4 HERE

Two correlated trait models M3 and M2 had acceptable fit and the indices reported in Table 4, but factor correlations were too high to consider the isolated factors as distinctive traits, or even distinctive facets of the same trait (the correlations ranged from .89 to .98). These findings, in addition to the presence of a strong general factor in all bifactor models as well as very high loading correlations between the single-factor solution and the G factors from the five different bifactor solutions, suggested that the BDI–II is essentially a unidimensional instrument, in spite of presence of some multidimensionality. Thus, we opted for the most parsimonious, unidimensional model – M1.

**Item Analysis of the BDI–II**

Given that the previous analyses supported unidimensionality of the BDI-II in the presence of some multidimensionality, we presented the factor loadings from the single factor model and the IRT graded response model item parameters (Table 5). It is important to note that the overall IRT model fit was good. A certain level of misfit involved the items #10, #1, and #12, however it could be considered negligible³.

³ Details available upon request from the first author.
All items had adequate slope parameter $a$, ranging from 1.38 (Loss of interest in Sex) to 2.89 (Sadness) indicating high to very high discrimination (Baker, 2001). The difficulty of item categories is distributed in $\theta$ to reflect the severity of depression. $\theta = 0$ reflects an average level of depression, while positive and negative values of theta indicate depression above and below average level, respectively. As can be seen in Table 5, category difficulty parameters ranged from -1.51 (Loss of pleasure) to .292 (Suicidal thoughts) for threshold between categories 0 and 1, from -.04 (Loss of interest in sex) to 1.47 (Suicidal thoughts) for threshold between categories 1 and 2, and from .58 (Loss of interest in sex) to 2.08 (Past failure) for threshold between categories 2 and 3. Overall, the lower the value of category threshold is, the lower level of depression is needed to endorse it. For example, it means that one has to be much more depressed to endorse category 1 on the Suicidal thoughts item then on the Loss of pleasure item.

Since GRM does not provide locations of items, we calculated a mean of category threshold parameters per item ($LI_{mean}$; Ali, Chang, & Anderson, 2015). According to $LI_{mean}$, locations of the majority of items was slightly above 0, ranging from -.19 to .63, with exception of Suicidal thoughts which had $LI_{mean}=1.26$. In terms of IRT, an item is most informative (precise) as a measure of a latent trait ($\theta$) at its location. Based on the values of the thresholds and $LI_{mean}$ reported in Table 5, one can conclude that the BDI–II is the most appropriate and, thus, most accurate for measuring depression, in the individuals with depression levels above average.

As can be seen from Figure 1, this Serbian adaptation of the BDI–II is the most informative in the $\theta$ range between -2 and 2 logits, but more (about 63% of information) in the individuals with levels of depression above average. Figure 1 also shows that the BDI-II has
outstanding precision in this clinical sample. Based on Figure 1, when test information is converted to classical reliability coefficient (Thissen, 2000), the BDI–II has reliability equal or above .90 (information over 10) in theta range -1.65–2.46 logits, and equal and above .95 (information over 20) in theta range -0.26–1.74 logits.

**Differential Item Functioning of the BDI–II**

In order to investigate measurement invariance of the BDI–II between genders, differential item functioning (DIF) analysis was performed. DIF can be manifested as difference in slopes (discrimination), difference in location parameters (difficulties), or both. Existence of DIF was tested using likelihood ratio test and comparing the fit of the baseline model with the fit of nested comparison models (see Meade, 2016).

Regarding difference in location parameters, gender-related DIF (using the anchor items approach proposed by Lopez Rivas, Stark, & Chernyshenko, 2009) was detected in the following items: #20, #18 and #21. According to standardized difference in expected scores (ESSD), which can be interpreted as Cohen’s d (Meade, 2016), females had slightly higher mean expected scores on these three items than males with the same level of depression. DIF effect size for item #20 was trivial (ESSD#20 = .196), small for item #18 (ESSD#18=.20) and moderate for item #21 (ESSD#21 =.71). However, maximum differences in the expected scores for items #18 and #21 were at below average levels of depression i.e., the part of the latent trait continuum which is not crucial for functioning of the BDI–II as a clinical instrument.

Item #10 (Crying) demonstrated a significant, small difference in slopes between genders

---

4 Details of the analyses are available from the first author upon request.
(ESSD_{#10} = -0.25). It means that the difference in expected scores between the two groups depended on the location of the theta continuum. As it can be seen in Figure 2, assuming the same level of depression, males with θ below average had higher expected scores on item #10 than females. On the other hand, males with θ above average had lower expected scores than females with the same level of depression. Slope is somewhat steeper in the female subsample, meaning that item #10 better differentiates subjects with different levels of depression in females, implying that crying is more related to the intensity of depression in females than males.

INSERT FIGURE 2 HERE

However, the existence of these differentially functioning items did not affect functioning of the test as a whole. Maximum difference in the expected test scores was less than 1 point (0.99) occurring at theta level of -0.83 (depression level below average). Effect size of this difference was trivial (ETSSD = 0.003). Signed test difference in the sample (STDS), which allows for cancelation of differences between items forming total score, was -0.049, while unsigned test difference in the sample (UTDS) was 2.06 points. The difference that exists between UTDS and STDS in the case of the Serbian adaptation of BDI–II shows that, although some differences in functioning at item level exist, they cancel each other out at the test level and do not affect the total score.

**Discussion**

Our results demonstrated unequivocally that the BDI–II, when used in a sample of clinically depressed individuals, is best regarded as a unidimensional instrument that measures the severity of depression. Even though two bifactor models had the best fit indices, additional model indices did not justify creation of subscales. For example, group factors were not well-
defined by their loadings. Only the G factor had acceptable values of EVC (.88), FD (.98), H (.97), and ωₜ (>.90), suggesting that this factor accounted for above 90% of the common variance in the bifactor models and that only the latent construct represented by the G factor is well-defined and replicable. Finally, there were very high correlations between the factor loadings in the unidimensional model and the loadings of the G factor in all bifactor models considered in this study, supporting our conclusion about unidimensionality of the Serbian version of the BDI–II. Hence, our results are in line with a number of previous studies that concluded that the BDI–II measures a single construct (depression severity), whereas clustering of the items tapping various aspects of this construct is not sufficient to form group or specific factors (e.g., Brouwer et al., 2013; Osman et al., 2008).

The results of the two-parameter IRT model suggested that all BDI–II items had high to very high discrimination (a parameter). The highest discrimination was demonstrated by the following items: Sadness, Loss of interest, Indecisiveness, and Concentration difficulty. It is noteworthy that these items are part of the A diagnostic criterion for major depressive disorder (APA, 2013), while the first two of them are the so-called main depression symptoms (one of them has to be fulfilled aside any of the other seven for diagnosis). From a somewhat different angle i.e., the network approach to psychopathology, recent studies in adult clinical samples have identified Sadness (Beard et al., 2016), Anhedonia (Bringmann, Lemmens, Huibers, Borsboom, & Tuerlinckx, 2015), and Fatigue (van Borkulo et al., 2015) as central symptoms. The central symptoms are those that have many and/or the strongest associations with other symptoms (Borsboom & Cramer, 2013). It can, then, be said that the most discriminative BDI–II items are, at the same time, the most central ones. Finally, these most discriminative items cover the whole range of depressive symptoms: affective, cognitive, and somatic-vegetative.

Based on the values of categories’ thresholds, which represent the point on the latent trait
of depression where the probability of responding above a certain category is .50, one can conclude that the thresholds were appropriate. Overall, the IRT threshold parameters for the BDI-II are spread over a reasonable range of the latent trait of depression severity (approximately from -2 to +2).

Based on the test information curve, in the clinical sample, the Serbian version of the BDI–II is the most informative in the medium range of depression severity (from -1.65 to 2.46 logits), but more (about 63% of information) in individuals with levels of depression above average. Since the test is primarily intended for clinical purposes, we can consider this to be optimal. In other words, the BDI–II is the most informative in the theta range where the majority of clinical subjects can be expected. Using the more familiar reliability terminology from the classical test theory, one can claim that the BDI–II has excellent reliability over a wide range of depression severities. It has excellent reliability (precision) just where it is needed the most.

Small to moderate DIF was found in three BDI–II items. Females tended to report more changes in appetite and interest in sex at all levels of depression (than equally depressed males). It is likely that such a pattern of responding reflects the fact that women, in general, have lower body image satisfaction than men (Hartmann, Rieger, & Vocks, 2019), and have a greater tendency to self-disclose sexual problems (Okur, van der Knaap, & Bogaerts, 2017).

A small non-uniform gender DIF was found for the item #10 (Crying). At the level of depression below average, males tended to disclose more crying than equally depressed females. The pattern was opposite at the higher level of depression severity, with females, on average, consistently opting for higher response options than their equally depressed male counterparts. Given that crying is a more common emotion-regulation strategy in females than in males (e.g., Vingerhoets & Scheirs, 2000), we can assume that for females who are at the lower end of the depression severity continuum it is difficult to discern whether crying represents an appropriate
emotional regulation strategy or a depressive symptom. Consequently, they don’t report that they cry more than otherwise. Additional information is that the crying item, in general, is more discriminative for females than males i.e., better reflects differences in depression severity in females than males. It seems that crying becomes a more prominent and discriminative symptom of depression in females when they become more depressed. Lower discrimination of crying in males is, on the other side, a consequence of the distribution of their answers. They opted mainly for extreme, “not crying”, categories. Thus, majority of their answers (47%) is in categories 0 – “boys don’t cry” (more than otherwise), and 3 – they want, but can’t cry (28%). This pattern of results supports a linear relation between crying and depression in females, but not males, for whom inability to cry is related to mild as well as severe levels of depression (see Vingerhoets, Rottenberg, Cevaal, & Nelson, 2007 for the review of depression-crying relationship models). Notwithstanding these differences that probably reflect different manifestations of depression between genders and/or their different help-seeking strategies, the total BDI–II score does not seem to be biased as an indicator of depression severity with respect to gender.

Limitations

First, sample in this study was a convenience sample, comprised of inpatients and outpatients of nine psychiatric hospitals in the Republic of Serbia. Since one of the aims of this paper was to examine the existence of gender-related measurement invariance, one can say that unequal gender groups (36.8% male, 63.2%) could be a problem. However, such distribution is in line with the epidemiological data and the higher prevalence of the depressive disorder in women (APA, 2013; Nolen-Hoeksema, 1987).

Although it was not small, sample size was another concern since it was below recommended 500 for GRM (Reise & Yu, 1990). Still, it was found that GRM slope parameters
estimates correlate with true parameters sufficiently (.85) in samples of \( N = 250 \) if the slope parameters are above .75 (factor loadings above .60) (Reise & Yu, 1990). Reise and Yu (1990) obtained these correlations in a 25-item test with five response categories. They found similar correlations between estimated and true difficulty parameters. Hence, we could consider the sample size acceptable having in mind the magnitude of factor loadings in our (selected) one factor model (≥ .63), and the clinical nature of the sample. It is still advisable to cross-validate findings in other samples.

References

Ali, U. S., Chang, H.-H., & Anderson, C. J. (2015). Location Indices for Ordinal Polytomous Items Based on Item Response Theory: IRT Location Indices for Ordinal Polytomous Items. *ETS Research Report Series, 2015*(2), 1–13. https://doi.org/10.1002/ets2.12065

Al-Turkait, F. A., & Ohaeri, J. U. (2010). Dimensional and Hierarchical Models of Depression Using the Beck Depression Inventory-II in an Arab College Student Sample. *BMC Psychiatry, 10*(1), Article 60. https://doi:10.1186/1471-244X-10-60

APA. (2013). *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.). Arlington, VA: American Psychological Association.

Baker, F. B. (2001). *The basics of item response theory* (2nd ed). College Park, Md.: ERIC Clearinghouse on Assessment and Evaluation.

Beard, C., Millner, A. J., Forgeard, M. J. C., Fried, E. I., Hsu, K. J., Treadway, M. T., … Björgvinsson, T. (2016). Network analysis of depression and anxiety symptom relationships in a psychiatric sample. *Psychological Medicine, 46*(16), 3359–3369. https://doi.org/10.1017/S0033291716002300

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *BDI-II, Beck depression inventory: Manual.*
San Antonio, Tex.: Psychological Corp.

Beck, A. T., Steer, R. A., & Brown, G. K. (2009). *Beck-Depressions-Inventar (BDI–II, dt. Version: M. Hautzinger, F. Keller & C. Kühner)* (2. Aufl.). Frankfurt: Pearson Assessment.

Beck, A. T., Steer, R. A., Brown, G. K., & Van der Does, W. (2002). *BDI-II-NL handleiding [BDI-II-Dutch manual]*. Lisse, the Netherlands: Psychological Corporation.

Beck, A. T., Ward, C., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An Inventory for Measuring Depression. *Archives of General Psychiatry, 4*(56), 561–571.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-Information Item Factor Analysis. *Applied Psychological Measurement, 12*(3), 261–280. https://doi.org/10.1177/014662168801200305

Bornovalova, M. A., Choate, A. M., Fatimah, H., Petersen, K. J., & Wiernik, B. M. (2020). Appropriate Use of Bifactor Analysis in Psychopathology Research: Appreciating Benefits and Limitations. *Biological Psychiatry, in press*, 1–15. https://doi.org/10.1016/j.biopsych.2020.01.013

Borsboom, D., & Cramer, A. O. (2013). Network Analysis: An Integrative Approach to the Structure of Psychopathology. *Annual Review of Clinical Psychology, 9*, 91–121.

Bringmann, L. F., Lemmens, L. H. J. M., Huibers, M. J. H., Borsboom, D., & Tuerlinckx, F. (2015). Revealing the dynamic network structure of the Beck Depression Inventory-II. *Psychological Medicine, 45*(4), 747–757. https://doi.org/10.1017/S0033291714001809

Brouwer, D., Meijer, R. R., & Zevalkink, J. (2013). On the Factor Structure of the Beck Depression Inventory–II: G Is the Key. *Psychological Assessment, 25*(1), 136–145. https://doi.org/10.1037/a0029228

Buckley, T. C., Parker, J. D., & Heggie, J. (2001). A Psychometric Evaluation of the BDI-II in
Treatment-Seeking Substance Abusers. *Journal of Substance Abuse Treatment*, 20(3), 197–204.

Bühler, J., Keller, F., & Läge, D. (2014). Activation as an Overlooked Factor in the BDI–II: A Factor Model Based on Core Symptoms and Qualitative Aspects of Depression. *Psychological Assessment*, 26(3), 970–979. https://doi.org/10.1037/a0036755

Chalmers, P. R. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, 48(6), 1–29. https://doi.org/10.18637/jss.v048.i06

de Miranda Azevedo, R., Roest, A. M., Carney, R. M., Denollet, J., Freedland, K. E., Grace, S. L., … others. (2016). A Bifactor Model of the Beck Depression Inventory and Its Association with Medical Prognosis After Myocardial Infarction. *Health Psychology*, 35(6), 614–624.

de Sá Junior, A. R., de Andrade, A. G., Andrade, L. H., Gorenstein, C., & Wang, Y.-P. (2018). Response Pattern of Depressive Symptoms Among College Students: What Lies Behind Items of the Beck Depression Inventory–II? *Journal of Affective Disorders*, 234, 124–130.

de Sá Junior, A. R., Liebel, G., Andrade, A. G., de Andrade, L. H., Gorenstein, C., & Wang, Y.-P. (2019). Can Gender and Age Impact on Response Pattern of Depressive Symptoms Among College Students? A Differential Item Functioning Analysis. *Frontiers in Psychiatry*, 10, Article 50. https://doi.org/10.3389/fpsyt.2019.00050

Dozois, D. J. A., Dobson, K. S., & Ahnberg, J. L. (1998). A Psychometric Evaluation of the Beck Depression Inventory–II. *Psychological Assessment*, 10(2), 83–89. https://doi.org/10.1037/1040-3590.10.2.83

Erford, B. T., Johnson, E., & Bardoshi, G. (2016). Meta-Analysis of the English Version of the
Beck Depression Inventory–Second Edition. *Measurement and Evaluation in Counseling and Development, 49*(1), 3–33.

Filzmoser, P., & Gschwandtner, M. (2018). *Mvoutlier: Multivariate Outlier Detection Based on Robust Methods.* R package version 2.0.9. Retrieved from https://CRAN.R-project.org/package=mvoutlier

Gorsuch, R. L. (1983). *Factor Analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Hancock, G. R., & Mueller, R. O. (2001). Rethinking Construct Reliability Within Latent Variable Systems. In R. Cudeck, K. G. Jöreskog, D. Sörbom, & S. Du Toit (Eds.), *Structural Equation Modeling: Present and Future; A festschrift in honor of Karl Joreskog; 2000* (pp. 195–216). Uppsala, Sweden: Scientific Software International.

Hartmann, A. S., Rieger, E., & Vocks, S. (2019). Sex and Gender Differences in Body Image. *Frontiers in Psychology, 10*, Article 1696. https://doi.org/10.3389/fpsyg.2019.01696

Hu, L., & Bentler, P. M. (1999). Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Kim, Y., Pilkonis, P. A., Frank, E., Thase, M. E., & Reynolds, C. F. (2002). Differential Functioning of the Beck Depression Inventory in Late-Life Patients: Use of Item Response Theory. *Psychology and Aging, 17*(3), 379–391.

Lopez Rivas, G. E., Stark, S., & Chernyshenko, O. S. (2009). The Effects of Referent Item Parameters on Differential Item Functioning Detection Using the Free Baseline Likelihood Ratio Test. *Applied Psychological Measurement, 33*(4), 251–265. https://doi.org/10.1177/0146621608321760

Markon, K. E. (2019). Bifactor and Hierarchical Models: Specification, Inference, and Interpretation. *Annual Review of Clinical Psychology, 15*, 51–69.
Meade, A. W. (2010). A Taxonomy of Effect Size Measures for the Differential Functioning of Items and Scales. *Journal of Applied Psychology, 95*(4), 728–743. https://doi.org/10.1037/a0018966

Meade, A. W. (2016). *Guide to IRT Invariance Using the MIRT Package in R*. North Carolina State University.

Maydeu-Olivares, A. (2013). Goodness-of-Fit Assessment of Item Response Theory Models. *Measurement: Interdisciplinary Research & Perspective, 11*(3), 71–101. https://doi.org/10.1080/15366367.2013.831680

Mihić, L., & Novović, Z. (2019). *Bekov inventar depresije—Drugo izdanje [Beck Depression Inventory—Second Edition]*. Belgrade, Serbia: Sinapsa.

Nolen-Hoeksema, S. (1987). Sex Differences in Unipolar Depression: Evidence and Theory. *Psychological Bulletin, 101*(2), 259–282.

Novović, Z., Mihić, L., Tovilović, S., Jovanović, V., & Biro, M. (2011). Psychometric Characteristics of the Beck Depression Inventory on a Serbian Student Sample. *Psihologija, 44*(3), 225–243. https://doi.org/10.2298/PSI1103225N

Okur, P., van der Knaap, L. M., & Bogaerts, S. (2017). A Quantitative Study on Gender Differences in Disclosing Child Sexual Abuse and Reasons for Nondisclosure. *Journal of Interpersonal Violence*, doi: 0886260517720732.

Osman, A., Barrios, F. X., Gutierrez, P. M., Williams, J. E., & Bailey, J. (2008). Psychometric Properties of the Beck Depression Inventory-II in Nonclinical Adolescent Samples. *Journal of Clinical Psychology, 64*(1), 83–102.

Patterson, A. L., Morasco, B. J., Fuller, B. E., Indest, D. W., Loftis, J. M., & Hauser, P. (2011). Screening for Depression in Patients with Hepatitis C Using the Beck Depression Inventory–II: Do Somatic Symptoms Compromise Validity? *General Hospital*
Quilty, L. C., Zhang, K. A., & Bagby, R. M. (2010). The Latent Symptom Structure of the Beck Depression Inventory–II in Outpatients with Major Depression. *Psychological Assessment, 22*(3), 603–608.

R Core Team. (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Reise, S. P., Kim, D. S., Mansolf, M., & Widaman, K. F. (2016). Is the Bifactor Model a Better Model or Is It Just Better at Modeling Implausible Responses? Application of Iteratively Reweighted Least Squares to the Rosenberg Self-Esteem Scale. *Multivariate Behavioral Research, 51*(6), 818–838. https://doi.org/10.1080/00273171.2016.1243461

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and Structural Coefficient Bias in Structural Equation Modeling: A Bifactor Perspective. *Educational and Psychological Measurement, 73*(1), 5–26. https://doi.org/10.1177/0013164412449831

Reise, S. P., Ventura, J., Keefe, R. S. E., Baade, L. E., Gold, J. M., Green, M. F., … Bilder, R. (2011). Bifactor and item response theory analyses of interviewer report scales of cognitive impairment in schizophrenia. *Psychological Assessment, 23*(1), 245–261. https://doi.org/10.1037/a0021501

Reise, S. P., & Yu, J. (1990). Parameter Recovery in the Graded Response Model Using MULTILOG. *Journal of Educational Measurement, 27*(2), 133–144. https://doi.org/10.1111/j.1745-3984.1990.tb00738.x

Revelle, W. (2009). *An introduction to psychometric theory with applications in R*. Springer.

Richter, J., Polak, T., & Eisemann, M. (2003). Depressive Mood and Personality in Terms of Temperament and Character Among the Normal Population and Depressive Inpatients.
Personality and Individual Differences, 35(4), 917–927.

Rizopoulos, D. (2006). ltm: An R Package for Latent Variable Modeling and Item Response Theory Analyses. *Journal of Statistical Software, 17*(5), 1-25. https://doi.org/10.18637/jss.v017.i05

Rodriguez-Gomez, J. R., Dávila-Martínez, M. G., & Collazo-Rodríguez, L. C. (2006). Factor Structure of the Beck Depression Inventory-(BDI-II) with Puerto Rican Elderly. *Puerto Rico Health Sciences Journal, 25*(2), 127–132.

Samejima, F. (1969). Estimation of Latent Ability Using a Response Pattern of Graded Scores. *Psychometrika Monograph Supplement, 34*(4, Pt. 2), 100–100.

Samejima, F. (1983). *Information Functions for General Model Developed for Differential Strategies in Cognitive Processes*. Tenessee University Knoxville.

Santor, D. A., Ramsay, J., & Zuroff, D. C. (1994). Nonparametric Item Analyses of the Beck Depression Inventory: Evaluating Gender Item Bias and Response Option Weights. *Psychological Assessment, 6*(3), 255–270. doi: 10.1037/1040-3590.6.3.255

Segal, D. L., Coolidge, F. L., Cahill, B. S., & O’Riley, A. A. (2008). Psychometric Properties of the Beck Depression Inventory—II (bBDI–II) Among Community-Dwelling Older Adults. *Behavior Modification, 32*(1), 3–20.

Seignourel, P. J., Green, C., & Schmitz, J. M. (2008). Factor Structure and Diagnostic Efficiency of the BDI-II in Treatment-Seeking Substance Users. *Drug and Alcohol Dependence, 93*(3), 271–278.

Subica, A. M., Fowler, J. C., Elhai, J. D., Frueh, B. C., Sharp, C., Kelly, E. L., & Allen, J. G. (2014). Factor Structure and Diagnostic Validity of the Beck Depression Inventory—II with Adult Clinical Inpatients: Comparison to a Gold-Standard Diagnostic Interview. *Psychological Assessment, 26*(4), 1106–1115. doi: 10.1037/a0036998
Subramaniam, G., Harrell, P., Huntley, E., & Tracy, M. (2009). Beck Depression Inventory for Depression Screening in Substance-Abusing Adolescents. *Journal of Substance Abuse Treatment, 37*(1), 25–31.

Thissen, David. (2000). Reliability and Measurement Precision. In H. Wainer, & N. J. Dorans (Eds.), *Computerized adaptive testing: A primer* (pp. 159–184). Mahwah, N.J: Lawrence Erlbaum Associates Publishers.

van Borkulo, C., Boschloo, L., Borsboom, D., Penninx, B. W. J. H., Waldorp, L. J., & Schoevers, R. A. (2015). Association of Symptom Network Structure With the Course of Depression. *JAMA Psychiatry, 72*(12), 1219–1226. https://doi.org/10.1001/jamapsychiatry.2015.2079

VanVoorhis, C. R. W., & Blumentritt, T. L. (2007). Psychometric Properties of the Beck Depression Inventory-II in a Clinically-Identified Sample of Mexican American Adolescents. *Journal of Child and Family Studies, 16*(6), 789–798.

Vingerhoets, A. J. M., & Scheirs, J. (2000). Sex Differences in Crying: Empirical Findings and Possible Explanations. In A. H. Fischer (Ed.), *Gender and emotion: Social psychological perspectives* (pp. 143–166). Cambridge: University Press.

Vingerhoets, A. J., Rottenberg, J., Cevaal, A., & Nelson, J. K. (2007). Is There a Relationship Between Depression and Crying? A Review. *Acta Psychiatrica Scandinavica, 115*(5), 340–351.

Ward, L. C. (2006). Comparison of Factor Structure Models for the Beck Depression Inventory—II. *Psychological Assessment, 18*(1), 81–88. https://doi.org/10.1037/1040-3590.18.1.81

World Health Organization. (1992). *International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD–10)*. Geneva: World Health Organization.
Dimenzionalnost i merna invarijantnost srpske verzije Bekovog inventara depresivnosti, BDI–II iz ugla teorije stavskog odgovora

Bojan B. Janičić, Ljiljana Mihić i Zdenka Novović

Odsek za psihologiju, Filozofski fakultet, Univerzitet u Novom Sadu, Srbija

Postoji debata oko dimenzionalnosti druge verzije Bekovog inventar depresivnosti (eng. the Beck’s Depression Inventory – II; BDI–II), adekvatnog načina njegovog skorovanja i merne invarijanosti . Pokušali smo da odgovorimo na ova pitanja koristeći Teoriju stavskog odgovora (TSO) na kliničkom uzorku od 288 pacijenata koristeći srpsku verziju ovog inventara. Proverili smo devet strukturnih modela koristeći konfirmativnu faktorsku analizu pune informativnosti (eng. Full Information Confirmatory Factor Analysis) i TSO model uređenih kategorija odgovora (eng. Graded Response Model). Zaključili smo da je BDI–II suštinski jednodimenzionalan. Svi ajtemi su visoko diskriminativni, a test je, na kliničkom uzorku, bio najinformativniji za umereni nivo depresije. Iako postoji malo do umereno diferencijalno funkcionisanje ajtema u odnosu na pol, ono ne utiče na ukupni skor. Shodno tome, ukupni skorovi ispitanika različitog pola na srpskoj verziji BDI–II kao mera nivoa depresije su međusobno uporedivi.

**Ključne reči:** depresija, BDI–II, Teorija stavskog odgovora, diferencijalno funkcionisanje ajtema, polne razlike
RECEIVED: 18.03.2020.

REVISION RECEIVED: 17.07.2020.

ACCEPTED: 25.08.2020.
### Tables

**Table 1**

*Descriptive Statistics*

| Item                  | Mean | SD  | Min | Max | Skewness | Kurtosis | SE  |
|-----------------------|------|-----|-----|-----|----------|----------|-----|
| 1. Sadness            | 1.17 | 1.02| 0   | 3   | .55      | -1.79    | .06 |
| 2. Pessimism          | 1.37 | 1.00| 0   | 3   | .26      | -1.01    | .06 |
| 3. Past failure       | 1.20 | .94 | 0   | 3   | .33      | -.82     | .06 |
| 4. Loss of pleasure   | 1.62 | .96 | 0   | 3   | .05      | -1.04    | .06 |
| 5. Guilty feelings    | 1.07 | .96 | 0   | 3   | .65      | -.49     | .06 |
| 6. Punishment feelings| 1.18 | 1.28| 0   | 3   | .48      | -1.49    | .08 |
| 7. Self-dislike       | 1.25 | .98 | 0   | 3   | .24      | -1.01    | .06 |
| 8. Self-criticalness  | 1.45 | 1.02| 0   | 3   | .17      | -.10     | .06 |
| 9. Suicidal thoughts  | .59  | .82 | 0   | 3   | 1.40     | 1.38     | .05 |
| 10. Crying            | 1.20 | 1.14| 0   | 3   | .42      | -1.27    | .07 |
| 11. Agitation         | 1.19 | .97 | 0   | 3   | .46      | -.77     | .06 |
| 12. Loss of interest  | 1.31 | .96 | 0   | 3   | .23      | -.91     | .06 |
| 13. Indecisiveness    | 1.30 | 1.02| 0   | 3   | .26      | -1.05    | .06 |
| 14. Worthlessness     | 1.11 | .99 | 0   | 3   | .54      | -.74     | .06 |
| 15. Loss of energy    | 1.50 | .93 | 0   | 3   | .01      | -.86     | .05 |
| 16. Changes in sleep  | 1.43 | .95 | 0   | 3   | .13      | -.91     | .06 |
| 17. Irritability      | 1.04 | .98 | 0   | 3   | .50      | -.86     | .06 |
| 18. Changes in appetite| 1.18 | .93 | 0   | 3   | .45      | -.62     | .05 |
| 19. Concentration difficulty | 1.35 | .90 | 0   | 3   | .22      | -.71     | .05 |
| 20. Tiredness or fatigue | 1.55 | 1.03| 0   | 3   | .02      | -1.17    | .06 |
| 21. Loss of interest in sex | 1.67 | 1.20| 0   | 3   | -.16     | -1.54    | .07 |
Table 2

**Detailed specifications of tested models**

| Model / Factors          | Type           | Items                               | Source                  |
|--------------------------|----------------|-------------------------------------|-------------------------|
| Model 1 (M1)             | Unidimensional | 1–21                                |                         |
| One factor               |                |                                     |                         |
| Model 2 (M2)             | Correlated     | 1–3, 5–9, 13, 14                    | Dozois et al. (1998)   |
| Cognitive-Affective (CA) | traits         | 4, 10–12, 15–21                     |                         |
| Somatic-Vegetative (SV)  |                |                                     |                         |
| Model 3 (M3)             | Correlated     | 3, 5–8, 13, 14                      | Beck et al. (2002)     |
| Cognitive (C)            | traits         | 10, 11, 15–21                       |                         |
| Somatic (S)              |                | 1, 2, 4, 9, 12                      |                         |
| Affective (A)            |                |                                     |                         |
| Model 4 (M4)             | Bifactor       | 1–21                                | Group factors          |
| General (G)              |                |                                     |                         |
| group Cognitive-Affective (gCA) |            | 1–3, 5–9, 13, 14                    | Group factors from Dozois et al. (1998) |
| group Somatic-Vegetative (gSV) |        | 4, 10–12, 15–21                     |                         |
| Model 5 (M5)             | Bifactor       | 1–21                                | Group factors          |
| General (G)              |                |                                     |                         |
| group Cognitive (gC)     |                | 3, 5–8, 13, 14                      | Beck et al. (2002)     |
| group Somatic (gS)       |                | 10, 11, 15–21                       |                         |
| group Affective (gA)     |                | 1, 2, 4, 9, 12                      |                         |
| Model 6 (M6)             | Bifactor       | 1–21                                | Ward (2006)            |
| General (G)              |                |                                     |                         |
| group Somatic (gS)       |                | 2, 3, 5–9, 14                       |                         |
| group Cognitive (gC)     |                | 15, 16, 18–20                       |                         |
| Model 7 (M7)             | Correlated     | 1–4, 6, 7, 9, 10, 12–15, 17, 21     | Novović et al. (2011) |
| Cognitive-Affective (CA) | traits         | 5, 8, 11, 16, 18–20                 |                         |
| Exhaustion (EX)          |                |                                     |                         |
| Model 8 (M8)             | Bifactor       | 1–21                                | Group factors          |
| General (G)              |                |                                     |                         |
| group Cognitive-Affective (gCA) |        | 1–4, 6, 7, 9, 10, 12–15, 17, 21     | Group factors from Novović et al. (2011) |
| group Exhaustion (gEX)   |                | 5, 8, 11, 16, 18–20                 |                         |
| Model 9 (M9)             | Bifactor       | 1–21                                | Bühler et al. (2014)   |
| General (G)              |                |                                     |                         |
| group Cognitive (gC)     |                | 2, 3, 5–9, 14                       |                         |
| group Somatic (gS)       |                | 10, 11, 16–21                       |                         |
| group Activation (gA)    |                | 6, 9–11, 15, 17, 19–21              |                         |
Table 3

*Fit indices of tested models*

| M2   | df  | p    | RMSEA | RMSEA 90%       | SRMSR | TLI    | CFI    | AICc  | SABIC  |
|------|-----|------|-------|-----------------|-------|--------|--------|-------|--------|
|      |     |      |       | RMSEA lower     |       |        |        |       |        |
|      |     |      |       | RMSEA upper     |       |        |        |       |        |
| M3   | 229 | .010 | .046  | .035            | .057  | .049   | .982   | .985  | 12790 | 12756.91 |
| M5   | 146 | .106 | .021  | .000            | .037  | .039   | .996   | .997  | 12792.47 | 12722.83 |
| M2   | 289 | .000 | .059  | .049            | .068  | .048   | .971   | .975  | 12802.23 | 12772.31 |
| M4   | 165 | .001 | .033  | .016            | .046  | .039   | .991   | .993  | 12802.37 | 12732.74 |
| M9   | 142 | .102 | .023  | .000            | .038  | .040   | .996   | .997  | 12812.20 | 12732.22 |
| M6   | 162 | .050 | .032  | .016            | .045  | .065   | .991   | .993  | 12826.09 | 12774.59 |
| M1   | 327 | .000 | .066  | .056            | .075  | .053   | .963   | .968  | 12847.70 | 12819.30 |
| M7   | 332 | .000 | .066  | .056            | .075  | .051   | .963   | .968  | 12855.20 | 12825.28 |
| M8   | 186 | .000 | .043  | .030            | .054  | .048   | .985   | .988  | 12867.71 | 12798.07 |

*Note.* Models are sorted according to AICc values.

M2 = Limited Information Goodness of Fit Statistic; RMSEA = Root Mean Square Error of Approximation; SRMSR = Standardized Root Mean Square Residual; TLI – Tucker-Lewis Index; CFI = Comparative Fit Index; AICc = Akaike information criterion (with correction for sample size); SABIC = Sample Adjusted Bayesian Information Criterion.
Table 4

Additional statistical indices for evaluating CFA models

|     | AL  | MML | MXL | NSL | ILF | PUC | ECV | ETV | \(\omega\) | \(\omega_h\) | \(\omega_R\) | FD | H |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----------|-----------|-----------|----|---|
| M1  | .76 | .61 | .86 | 21  | 21  | 1.00| .57 | .96 | .98       | .97       | .97       | .97| .97|
| M2  | CAF | .77 | .65 | .86 | 10  | 10  | .47 | .28 | .93       | .97       | .94       | .97| .95|
|     | SVG | .78 | .65 | .86 | 11  | 11  | .53 | .31 | .94       | .97       | .97       | .97| .92|
| M3  | CF  | .77 | .67 | .86 | 7   | 7   | .33 | .19 | .91       | .96       | .92       | .97| .92|
|     | SF  | .76 | .65 | .85 | 9   | 9   | .41 | .24 | .92       | .97       | .93       | .97| .93|
|     | AF  | .82 | .73 | .87 | 5   | 5   | .27 | .16 | .91       | .96       | .94       | .96| .91|
| M4  | G   | .72 | .57 | .86 | 21  | 21  | .52 | .86 | .98       | .92       | .93       | .97| .96|
|     | CAF | .28 | -0.00| .69 | 3   | 10  | .09 | .05 | .94       | .12       | .12       | .86| .62|
|     | SVG | .21 | -0.05| .38 | 4   | 11  | .05 | .03 | .94       | .07       | .07       | .74| .44|
| M5  | G   | .75 | .57 | .87 | 21  | 21  | .68 | .88 | .99       | .94       | .95       | .98| .97|
|     | CF  | .37 | .19 | .61 | 4   | 7   | .08 | .05 | .92       | .20       | .22       | .85| .58|
|     | SF  | .12 | -1.13| .36 | 1   | 9   | .03 | .02 | .93       | .02       | .02       | .67| .29|
|     | AF  | .09 | -1.13| .41 | 1   | 5   | .02 | .01 | .92       | .01       | .01       | .78| .20|
| M6  | G   | .70 | .52 | .82 | 21  | 21  | .82 | .88 | .98       | .94       | .96       | .98| .96|
|     | CF  | .35 | .04 | .59 | 5   | 8   | .10 | .05 | .90       | .20       | .22       | .83| .59|
|     | SF  | -.12| -0.49| .16 | 1   | 5   | .03 | .02 | .88       | .02       | .02       | .67| .28|
| M7  | CAF | .77 | .61 | .86 | 14  | 14  | .68 | .38 | .95       | .98       | .96       | .95| .91|
|     | EX  | .75 | .67 | .86 | 7   | 7   | .32 | .18 | .89       | .95       | .91       |     |    |
| M8  | G   | .72 | .51 | .85 | 21  | 21  | .47 | .87 | .51       | .98       | .92       | .93       | .98| .96|
|     | CAF | .23 | .01 | .49 | 4   | 14  | .08 | .05 | .95       | .09       | .09       | .76| .54|
|     | EX  | -.15| -.57| .17 | 2   | 7   | .05 | .03 | .90       | .03       | .03       | .82| .49|
| M9  | G   | .75 | .59 | .86 | 21  | 21  | .56 | .88 | .56       | .99       | .93       | .94       | .98| .97|
|     | CF  | .28 | -0.05| .56 | 4   | 8   | .07 | .04 | .92       | .12       | .13       | .82| .53|
|     | SF  | .19 | -0.07| .31 | 0   | 8   | .03 | .02 | .91       | .06       | .06       | .67| .32|
|     | AC  | .02 | -0.24| .30 | 0   | 9   | .02 | .01 | .93       | .00       | .00       | .69| .24|

Note. AL = average loading using Fisher’s z transformation; MML = minimal loading; MXL = maximal loading; NSL = number of loadings > .32; ILF = number of items that load on factor; PUC = proportion of uncontaminated correlations; ECV = explained common variance; ETV = explained total variance; \(\omega\) = Omega reliability coefficient for general and group factors; \(\omega_h\) = Omega hierarchical reliability coefficient for general and group factors; \(\omega_R\) = Relative Omega coefficient \((\omega_h/\omega)\); FD = Factor determinacy; H = index of construct replicability

Factors: G = General; CAF = Cognitive-Affective; SVG = Somatic-Vegetative; CF = Cognitive; AF = Affective; SF = Somatic; EX = Exhaustion; AC = Activation.
Table 5

*Table 5 Single factor model loadings and IRT graded response model item parameters*

| Item                  | λ   | $h^2$ | a    | $b_1$ | $b_2$ | $b_3$ | $LI_{mean}$ |
|-----------------------|-----|-------|------|-------|-------|-------|-------------|
| Sadness              | .86 | .74   | 2.89 | -0.61 | 0.67  | 1.18  | 0.41        |
| Pessimism            | .72 | .52   | 1.77 | -1.13 | 0.33  | 1.29  | 0.16        |
| Past failure         | .61 | .37   | 1.32 | -1.02 | 0.60  | 2.08  | 0.55        |
| Loss of pleasure     | .82 | .67   | 2.41 | -1.52 | 0.04  | 0.93  | -0.18       |
| Guilty feelings      | .71 | .51   | 1.73 | -0.67 | 0.87  | 1.68  | 0.63        |
| Punishment feelings  | .64 | .41   | 1.41 | -0.14 | 0.63  | 0.92  | 0.47        |
| Self-dislike         | .78 | .60   | 2.09 | -0.74 | 0.43  | 1.52  | 0.40        |
| Self-criticalness    | .67 | .45   | 1.53 | -1.29 | 0.26  | 1.23  | 0.07        |
| Suicidal thoughts    | .80 | .64   | 2.24 | -0.29 | 1.47  | 2.01  | 1.26        |
| Crying               | .66 | .44   | 1.50 | -0.43 | 0.68  | 1.34  | 0.53        |
| Agitation            | .74 | .55   | 1.86 | -0.83 | 0.68  | 1.53  | 0.46        |
| Loss of interest     | .85 | .72   | 2.73 | -0.90 | 0.31  | 1.34  | 0.25        |
| Indecisiveness       | .85 | .73   | 2.78 | -0.75 | 0.37  | 1.21  | 0.28        |
| Worthlessness        | .79 | .62   | 2.17 | -0.58 | 0.69  | 1.48  | 0.53        |
| Loss of energy       | .84 | .70   | 2.58 | -1.28 | 0.08  | 1.24  | 0.02        |
| Changes in sleep     | .69 | .47   | 1.61 | -1.39 | 0.24  | 1.52  | 0.12        |
| Irritability         | .74 | .55   | 1.88 | -0.45 | 0.65  | 1.78  | 0.66        |
| Changes in appetite  | .68 | .46   | 1.56 | -1.01 | 0.71  | 1.84  | 0.52        |
| Concentration difficulty | .85 | .72   | 2.75 | -1.15 | 0.33  | 1.39  | 0.19        |
| Tiredness or fatigue | .82 | .67   | 2.45 | -1.17 | 0.07  | 0.92  | -0.06       |
| Loss of interest in sex | .63 | .40   | 1.38 | -1.10 | 0.05  | 0.58  | -0.19       |

*Note.* $\lambda =$ Standardized factor loadings; $h^2 =$ Communality; $a =$ Slope (discrimination parameter); $b_1$, $b_2$, and $b_3 =$ Location/difficulty parameters ($\theta$ level needed to move from one response category to another [response categories are 0, 1, 2 and 3]); $LI_{mean} =$ item location index (mean);
Figure 1. **Test information curve for the Serbian version of the BDI–II**

*Gray area marks theta interval in which test reliability is > .90, and pattern area marks theta interval in which reliability is > .95*
Figure 2. Expected scores for gender groups on item #10.