Manuscript 13.8. 2019

The General Health Questionnaire (GHQ-12), Beck Depression Inventory (BDI-6), and Mental Health Index (MHI-5): psychometric and predictive properties in a Finnish population-based sample

Marko Elovanio, PhD[a,b]

Christian Hakulinen, PhD[a]

Laura Pulkki-Råback, PhD[a]

Anna-Mari Aalto, PhD[b]

Marianna Virtanen, PhD[c]

Timo Partonen, MD, PhD[b]

Jaana Suvisaari, MD, PhD[b]

[a]Department of Psychology and Logopedics, Faculty of Medicine, University of Helsinki, Finland;

[b]National Institute for Health and Welfare, Finland;

[c]School of Educational Sciences and Psychology, University of Eastern Finland, Joensuu, Finland

Correspondence: Marko Elovainio, University of Helsinki, P.O. Box 9, 00014, Helsinki, Finland, Phone: +358 50 3020621, email: marko.elovainio@helsinki.fi

Word count for abstract / main text: 250 / 2553
Number of tables and figures: 3 + 8
Keywords: Distress; depression; population; psychometrics; measuring

Word count for abstract / main text: 170 / 2553
Number of tables and figures: 1 + 3
Keywords: Distress; depression; population; psychometrics; measuring

The authors declare no conflict of interests.

Abstract

The short versions of the General Health Questionnaire (GHQ-12), Beck's Depression Inventory (BDI-6), and Mental Health Index (MHI-5) are all valid and reliable measures of general psychological distress, depressive symptoms, and anxiety. We tested the psychometric properties of the scales, their overlap, and their ability to predict mental health service use using both regression and machine learning (ML, random forest) approaches. Data were from the population-based

FinHealth-2017 Study of adults (N=4270) with data on all of the evaluated instruments. Constructive validity, internal consistency, invariance, and optimal cut-off points in predicting mental health services were tested. Constructive validity was acceptable and all instruments measured their own distinct phenomenon. Some of the item scoring in BDI-6 was not optimal, and the sensitivity and specificity of all scales were relatively weak in predicting service use. Small gender differences emerged in optimal cut-off points. ML did not improve model predictions. GHQ-12, BDI-6, and MHI-5 may be interpreted to measure different constructs of psychological health symptoms, but are not particularly useful predictors of service use.

Introduction

Measuring symptoms of depression, anxiety and general psychological distress using valid instruments is of great importance in healthcare settings. Commonly used instruments include the 12-item version of the General Health Questionnaire (GHQ-12) (Goldberg et al., 1997), the short Mental Health Index (MHI-5) questionnaire from the the Medical Outcomes Study Short-Form 36-Item Health Survey (SF-36) (Berwick et al., 1991; Ware et al., 1995), and various forms of the Beck Depression Inventory (BDI) (Beck, 1967; Beck and Beck, 1972). All of these scales were originally validated in relatively small clinical (Johnstone and Goldberg, 1976) or community samples (Banks, 1983).

The GHQ was developed as a screening instrument of psychological distress to be used in general practice, but it has also commonly been used in epidemiological survey studies. Of the different versions, the 12-item version, GHQ-12, has been the most frequently used. Older studies have shown GHQ-12 to have acceptable psychometric properties (McCabe et al., 1996; Schmitz et al., 1999). The original 21-item version of BDI is probably the most widely used survey and screening instrument of depressive symptoms and depression with good psychometric properties (Wang and Gorenstein, 2013). The shorter versions of BDI, including the six-item version used here, have been suggested to have comparable psychometric properties to the original instrument, but only one study has actually tested it (Aalto et al., 2012). Similarly, MHI-5 is a short measure of mental health dimension originally incorporated into a longer instrument (RAND-36) tapping both negative and positive aspects of mood, and evidence of its validity as an independent measure of mental health has been presented (Hoeymans et al., 2004; Kelly et al., 2008; McCabe et al., 1996; Rivera-Riquelme et al., 2019; Rumpf et al., 2001; Strand et al., 2003; Thorsen et al., 2013; Trainor et al., 2013) .

All of the scales have separately been shown to be good or reasonably valid and reliable instruments for measuring mental health symptoms, although previous studies have found differences in the ability of the scales to differentiate between depressive and other psychiatric disorders among genders (Aalto et al., 2012; Schmitz et al., 1999). Little is also known about the overlap between the scales, the item-level associations across measures, the structural invariance across genders (whether the constructs are similar in men and women), and the most important items that potentially predict the use of health mental services. Moreover, no earlier studies have thoroughly tested the psychometric properties of these scales at the item level.

We tested the structural validity and item-level associations across the GHQ-12, BDI-6, and MHI-5 and assessed the capacity of these instruments to predict mental health service use in a large nationally representative sample of the Finnish population. We used exploratory and confirmatory factor analyses and item-response modeling in evaluating the basic psychometric properties of the scales and the items included. We also tested whether the scales could predict mental health service use separately in men and women. Furthermore, we used a machine learning (ML) approach (random forest) to evaluate the predictability of the items included in all of these scales. The main aim was to explore whether these scales are interchangeable or whether they measure reasonably unique dimensions of mental health outcomes.

Methods

The participants were from the population-based FinHealth-2017 health examination study carried out in 50 localities in 2017, with the objective of evaluating 10 000 randomly selected persons aged over 18 years in Finland. The study consists of a physical examination and questionnaires (Borodulin and Sääksjärvi, 2019). In total, 7 055 people participated, yielding a response rate of 69%. Of the respondents, 56% were women and the mean age was 53.9 years (Table 1).

Measures

We measured psychological distress and depressive symptoms (or the lack of them) using three difference scales: the 12-item version of the General Health Questionnaire (GHQ-12, (Goldberg et al., 1997), the Mental Health Inventory (MHI-5) derived from the SF-36 scale (McHorney and Ware, 1995), and the short version of the Beck Depression Inventory (BDI-6) (Aalto et al., 2012). The GHQ-12 includes 12 questions assessing symptoms related to psychological distress and general functioning, e.g. ability to face problems and make decisions. All items have a 4-point scoring system ranging from a "*better/healthier than normal*" option, through to "*same as usual*", "*worse/more than usual*' or "*much worse/more than usual*". These are further scored using a 0-0-1-1- scoring such that *"better"* and *"usual"* responses are scored as 0, and *"worse"* and *"much* *worse"* responses are scored as 1. The responses to individual items are summed to yield a total score varying from 0 to 12. The total scale can be found from https://www.gl-assessment.co.uk/products/general-health-questionnaire-ghq.

The MHI-5 includes five questions covering the past four weeks: (1) Have you been a very nervous person?, (2) Have you felt so down in the dumps that nothing could cheer you up?, (3) Have you felt calm and peaceful?, (4) Have you felt downhearted and blue?, and (5) Have you been a happy person? All items have a 6-point scoring system ranging from "*all of the time*" to "*none of the time*" that are further scaled from 0 to 100.

The 6-item version of the Beck Depression Inventory (BDI-6) (Aalto et al., 2012), was derived from the original 21-item BDI (Beck et al., 1961) and includes the following items: depressed mood, pessimism, dissatisfaction, guilt, self-dislike and indecisiveness. The complete scale can be found in the appendix. The scoring of each of the BDI-6 questions was from 1 to 5 and was transformed into a BDI original scoring from 0 to 3 (1=0, 2=1, 3=2, 4=2, 5=3) to yield a total score varying from 0 to 18.

As criteria variables for validity, we used the following question: "Have you used any health services during the past 12 months due to mental health problems?"

Statistical analyses

The psychometric properties of the three scales were analyzed in the following steps: First, differences in symptom reporting between men and women were tested using ANOVA separately for each scale, and the preliminary structural analyses were conducted by calculating bivariate correlations between study variables. Second, the factorial validity of the original scales was tested in all participants and separately among men and women using exploratory factor analyses (Oblimin rotation; (Maharee-Lawler et al., 2010), followed by confirmatory factor analyses (CFA) (Jöreskog, 1993). We tested the factor structure and number of dimensions using exploratory factor analyses with eigenvalue 1 and loading structure as a criterion for the appropriate number of factors. Goodness of fit of the CFA models was evaluated based on Chi-square test (X^2), root mean squared error of approximation (RMSEA), the comparative fit index (CFI), the Tucker–Lewis index (TLI), and Akaike's information criterion (AIC) (Kelloway, 1998). Testing the final structure was done in three steps: (A) a one-factor model was estimated where all remaining items loaded on the same underlying dimension (null model), (B) a model representing the original theoretical model was estimated, and (C) the structural invariance was tested between men and women as strong scalar invariance with factor loadings and intercepts constrained to be similar.

Third, we tested the item properties (difficulty and discrimination), the term referring to a subject being unlikely to answer in the keyed direction (a patient is unlikely to endorse), using item response (IRT) modeling (DeAyala, 2009) with Generalized Partial Credit Models (GPCMs). GPCMs were conducted using the original scaling versions of the scales and also scaling MHI-5 into the same number of options as the others. Fourth, concurrent validity was assessed using self-reported service use for mental health problems during the last 12 months. Receiver operating characteristics (ROC) curve analyses were conducted with results interpreted as acceptable if the area under the curve (AUC) values were between 0.70 and 0.79, and good if they were ≥0.80. We estimated the optimal cut-off points for each scale using maximizing sensitivity and specificity separately in men and women.

Fifth, we used regression trees (Strobl et al., 2009) partitioning of the covariate space (of all predictor variables) to generate a final set of predictor variables and cut-off values within these predictors to find non-overlapping groups of subjects with similar values for a selected response variable. Group membership can then be determined by running through the hierarchy of nodes, which are those predictors that best explain heterogeneity in the cohort. Regression tree analyses were conducted using a random forest algorithm with the R-package "Caret". Random forests are sets of independently grown regression trees (bootstrapped), where each tree is weighted in order to calculate each predictor's importance.

R (version 3.5.1) were used for the statistical analyses.

Results

As expected, none of the items were normally distributed (D'Agostino's K^2 test p-values all <0.001). Women reported more symptoms or lower mental health than men irrespective of the scale (Table 1). The inner consistency (Cronbach's alphas) was good in all scales (0.92 for GHQ-12, 0.94 for BDI-6, and 0.89 for MHI-5). Correlation analyses among all variables suggested three to four structures in all subjects and in both men and women.

Exploratory factor analyses supported the choice of three factors by three methods (Optimal Coordinates, Parallel Analysis, Eigenvalues (Kaiser Criterion)) in all subjects and in both men and women. The factors solutions were almost identical in all subjects and in both genders (Figure 1). In men and women, the confirmatory factor analyses with three latent constructs provided acceptable fit to the data (CFI = 0.92, TLI = 0.91, NFI = 0.91, RMSEA = 0.063, SRMR = 0.036), and the three-factor solution was significantly better than the one-factor model (Chi-square diff (3) = 1085, p<0.001). The fit indices were nearly identical in men and women (CFI = 0.92/0.91, TLI = 0.91/0.90, NFI = 0.91/0.90, RMSEA =0.063/0.065, SRMR = 0.036/0.037), although the invariance test did not support strong invariance between men and women (Chi-square diff (20) = 31, p<0.05) between configural loading invariance.

GPCM also showed that all items loaded relatively well with their corresponding latent constructs (all loadings were over 0.51). All items showed acceptable discriminability and item categories showed logical difficulty (increasing difficulty with increasing symptoms). The information test showed that most of the information in BDI-6 and MHI-5 was in the relatively severe end of the scales, but in GHQ-12 the information curve had two peaks (Figure 2). Item trace lines suggest that the scoring of the items work well, except in two BDI-6 items, which seemed to be three-step rather than four-step items (questions about pessimism and dissatisfaction) (Figure 3).

These findings are supported by the fact that in both items the respondents had chosen option three more often than option two (less severe).

When using mental health service use as a criterion, the AUC of the scales in all participants was 0.76 for GHQ-12, 0.77 for BDI-6, and 0.79 for MHI-5. The optimal cut-off point for GHQ-12 was 3 (sensitivity = 0.62 / specificity = 0.79) for all, 4 for women (sensitivity = 0.57 / specificity = 0.82, AUC = 0.75), and 3 for men (sensitivity = 0.61 / specificity = 0.83, AUC = 0.77). For BDI-6, the optimal cut-off point for all was 2 (sensitivity = 0.74 / specificity = 0.77). The optimal cut-off point for women was 2 (sensitivity = 0.72 / specificity = 0.71, AUC = 0.76) and for men 2 (sensitivity = 0.66, specificity = 0.79, AUC = 0.78). For MHI-5, the optimal cut-off point was 72 (sensitivity = 0.84 / specificity = 0.59) in all and 76 in women (sensitivity = 0.76 / specificity = 0.65, AUC = 0.78), but in men the optimal cut-off point was 72 (sensitivity = 0.83 / specificity = 0.63, AUC = 0.81) (Figure 5). The cut-off point of 3/4 in GHQ-12, 3.4/3.5 in BDI -6, and 71/72 in MHI-5 showed very similar prevalence distributions in both men and women.

The random forest algorithm did not markedly improve the predictive ability of the items in the three scales. When using the mental health service use as the predicted outcome, the model showed 0.77 accuracy (95% CI 0.74 to 0.79), with 0.93 positive prediction value and only 0.20 negative prediction value. The most important items in predicting service use were “Have you felt calm and peaceful? (mh3)”, “Feeling sad? (bdi1)”, “Have you been very nervous person? (mh1)”, “Feeling guilty? (bdi4)”, “Have you felt so down in the dumps that nothing could cheer you up? (mh2)”, and “Have you felt downhearted and blue? (mh4)”.

Discussion

This study investigated the degree to which three mental health scales differ in their item content and predictive validity. Multiple instruments exist for screening psychological distress or symptoms of depression and anxiety in the general population, but their performance in relation to each other is unclear. The most common practice is to assess mental health or mental health problems with one particular scale and then draw conclusions relying on the assumption that the scales are interchangeable.

Earlier studies have found relatively subtle differences between different screening instruments (Aalto et al., 2012; Kelly et al., 2008; McCabe et al., 1996). Our results clearly support this conclusion. All of the included instruments (GHQ-12, BDI-6, and MHI-5) are commonly used to measure slightly different dimensions of mental health (GHQ, general psychological distress; BDI, mood problems; and MHI, general mental health and well-being), but all have a relatively wide focus. In our study, all of the scales showed acceptable psychometric properties with good internal consistency and independent factor structure and item characteristics. Only two of the BDI items had questionable item characteristics, which may well be caused by our use of the original BDI, which had to be rescaled from a 5- to a 4-point scale to obtain comparable cut-off points.

The exploratory and confirmatory factor analyses and item response models suggested that all of the items loaded well on their corresponding latent variables, supporting the interpretation that there was no considerable overlap between the scales and that all scales measure a separate construct or dimension of mental health (problems). None of the scales were particularly successful in predicting mental health service use in men or in women. Our analyses suggested some differences in the best cut-off points of the scales in relation to mental health service use, but in some the cut-off points were higher in women (BDI) and in some in men (GHQ). Using the average cut-off points divided the population into very similar case-non-case groups, suggesting similar prevalences of mental health problems in the population and also indicating that the same cut-off points can be used for men and women.

We studied the scales against mental health service use, which is, of course, not the same as having a mental disorder. Earlier studies have found that men are less likely to seek treatment for common mental disorders (Roberts et al., 2018; ten Have et al., 2013; Wang et al., 2007a; Wang et al., 2007b), and this gender gap has been particularly large for mood disorders (Kovess-Masfety et al., 2014). However, previous Finnish studies have not found significant gender differences in treatment seeking for depressive and anxiety disorders, although differences have emerged in the type of treatment received (Hamalainen et al., 2008; Kasteenpohja et al., 2015, 2016).

Using all items included in all scales and the ML approach (random forest) did not markedly improve the predictions. The most important features based on the random forest algorithm were those from the MHI-5 and BDI-6. However, because these items are highly correlated, they are not the best possible candidates in random forest or any other ML algorithms, and thus, the order of importance may be questioned. Our being constrained to a specific set of symptom data for determination of multivariate pathways may be one of the most serious limitations of the study; adding variables from other areas may have resulted in other solutions. We were, however, interested in how well these widely used instruments predict service use. We were able to use a large population-based random sample of people considered generally healthy, and all of the cut-off points suggested by the statistical models provided similar prevalence figures of mental health problems, which may not, however, be interpreted as prevalence estimates of clinically meaningful disorders.

The scales tested here, BDI-6, GHQ-12, and MHI-5, seem to measure slightly different aspects of mental health or mental health problems, but they should not be used solely in evaluating or predicting the need for mental health services or anxiolytic/antidepressant medication use in community samples. The psychometric properties of these scales were acceptable, and they may be useful in associative studies (examining associations between mental health and various outcomes). The overlap between scales was low, and thus, our results suggest that the common practice of using one particular scale in mental health research may lead to questionable results. The lack of overlap may be due to multiple reasons. It may reflect the differences in theoretical backgrounds, clinical opinions, or original purposes of the scales. The sum scores of the scales were relatively strongly correlated with each other, but this does not imply that the scales measure the same construct.

In conclusion, measuring mental health problems is important in a wide variety of health research, including epidemiological studies, studies evaluating health service performance, and intervention studies promoting health. Although there is a need for valid short-form instruments for measuring mental health in survey studies, our results suggest that until we better understand scale performance, a more conservative approach includes the use of multiple instruments, bearing in mind their theoretical contents. If different instruments capture different aspects of the mental health concept, there is the risk that the selection of a particular scale for a study may severely bias the results.

Funding: ME and CH were supported by the Academy of Finland (329224 (ME) / 310591(CH)).

References

Aalto, A.M., Elovainio, M., Kivimaki, M., Uutela, A., Pirkola, S., 2012. The Beck Depression Inventory and General Health Questionnaire as measures of depression in the general population: a validation study using the Composite International Diagnostic Interview as the gold standard. Psychiatry Res 197 (1-2), 163-171.

Banks, M.H., 1983. Validation of the General Health Questionnaire in a young community sample. Psychol Med 13 (2), 349-353.

Beck, A., 1967. Depression. Clinical, experimental and theoretical aspects. New York, Harper & Row Publishers, 1967. Harper & Row Publishers, New York.

Beck, A., Beck, R., 1972. Screening depressed patients in family practice. A rapid technic. Postgrad Med. 52, 81-85.

Beck, A.T., Ward, C.H., Mendelson, M., Mock, J., Erbaugh, J., 1961. An inventory for measuring depression. Arch Gen Psychiatry 4, 561-571.

Berwick, D.M., Murphy, J.M., Goldman, P.A., Ware, J.E., Jr., Barsky, A.J., Weinstein, M.C., 1991. Performance of a five-item mental health screening test. Med Care 29 (2), 169-176.

Borodulin, K., Sääksjärvi, K., 2019. FinHealth 2017 Study – Methods., Finnish Institute for Health and Welfare. Report 17/2019. . THL, Helsinki, p. 132.

DeAyala, R., 2009. The Theory and Practice of Item Response Theory. . The Guilford Press New york.

Goldberg, D.P., Gater, R., Sartorius, N., Ustun, T.B., Piccinelli, M., Gureje, O., Rutter, C., 1997. The validity of two versions of the GHQ in the WHO study of mental illness in general health care. Psychological Medicine 27 (1), 191-197.

Hamalainen, J., Isometsa, E., Sihvo, S., Pirkola, S., Kiviruusu, O., 2008. Use of health services for major depressive and anxiety disorders in Finland. Depress Anxiety 25 (1), 27-37.

Hoeymans, N., Garssen, A.A., Westert, G.P., Verhaak, P.F., 2004. Measuring mental health of the Dutch population: a comparison of the GHQ-12 and the MHI-5. Health Qual Life Outcomes 2, 23.

Johnstone, A., Goldberg, D., 1976. Psychiatric screening in general practice. A controlled trial. Lancet 1 (7960), 605-608.

Jöreskog, K.G., 1993. Testing structural equation models, in: Bollen, K.A., Long, J.S. (Eds.), Testing structural equation models. Sage, Newbury Park, pp. 294-315.

Kasteenpohja, T., Marttunen, M., Aalto-Setala, T., Perala, J., Saarni, S.I., Suvisaari, J., 2015. Treatment received and treatment adequacy of depressive disorders among young adults in Finland. BMC Psychiatry 15, 47.

Kasteenpohja, T., Marttunen, M., Aalto-Setala, T., Perala, J., Saarni, S.I., Suvisaari, J., 2016. Treatment adequacy of anxiety disorders among young adults in Finland. BMC Psychiatry 16, 63.

Kelloway, E.K., 1998. Using LISREL for structural equation modeling: A researcher's guide. SAGE, Thousand Oaks, CA.

Kelly, M.J., Dunstan, F.D., Lloyd, K., Fone, D.L., 2008. Evaluating cutpoints for the MHI-5 and MCS using the GHQ-12: a comparison of five different methods. BMC Psychiatry 8, 10.

Kovess-Masfety, V., Boyd, A., van de Velde, S., de Graaf, R., Vilagut, G., Haro, J.M., Florescu, S., O'Neill, S., Weinberg, L., Alonso, J., investigators, E.-W., 2014. Are there gender differences in service use for mental disorders across countries in the European Union? Results from the EU-World Mental Health survey. J Epidemiol Community Health 68 (7), 649-656.

Maharee-Lawler, S., Rodwell, J., Noblet, A.J., 2010. A step toward a common measure of organizational justice. Psychological Reports 106, 407-418.
McCabe, C.J., Thomas, K.J., Brazier, J.E., Coleman, P., 1996. Measuring the mental health status of a population: a comparison of the GHQ-12 and the SF-36 (MHI-5). Br J Psychiatry 169 (4), 516-521.
McHorney, C.A., Ware, J.E., Jr., 1995. Construction and validation of an alternate form general mental health scale for the Medical Outcomes Study Short-Form 36-Item Health Survey. Med Care 33 (1), 15-28.
Rivera-Riquelme, M., Piqueras, J.A., Cuijpers, P., 2019. The Revised Mental Health Inventory-5 (MHI-5) as an ultra-brief screening measure of bidimensional mental health in children and adolescents. Psychiatry Research 274, 247-253.
Roberts, T., Miguel Esponda, G., Krupchanka, D., Shidhaye, R., Patel, V., Rathod, S., 2018. Factors associated with health service utilisation for common mental disorders: a systematic review. BMC Psychiatry 18 (1), 262.
Rumpf, H.J., Meyer, C., Hapke, U., John, U., 2001. Screening for mental health: validity of the MHI-5 using DSM-IV Axis I psychiatric disorders as gold standard. Psychiatry Res 105 (3), 243-253.
Schmitz, N., Kruse, J., Tress, W., 1999. Psychometric properties of the General Health Questionnaire (GHQ-12) in a German primary care sample. Acta Psychiatrica Scandinavica 100 (6), 462-468.
Strand, B.H., Dalgard, O.S., Tambs, K., Rognerud, M., 2003. Measuring the mental health status of the Norwegian population: a comparison of the instruments SCL-25, SCL-10, SCL-5 and MHI-5 (SF-36). Nord J Psychiatry 57 (2), 113-118.
Strobl, C., Malley, J., Tutz, G., 2009. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychol Methods 14 (4), 323-348.
ten Have, M., de Graaf, R., van Dorsselaer, S., Beekman, A., 2013. Lifetime treatment contact and delay in treatment seeking after first onset of a mental disorder. Psychiatr Serv 64 (10), 981-989.
Thorsen, S.V., Rugulies, R., Hjarsbech, P.U., Bjorner, J.B., 2013. The predictive value of mental health for long-term sickness absence: the Major Depression Inventory (MDI) and the Mental Health Inventory (MHI-5) compared. BMC Med Res Methodol 13, 115.
Trainor, K., Mallett, J., Rushe, T., 2013. Age related differences in mental health scale scores and depression diagnosis: adult responses to the CIDI-SF and MHI-5. J Affect Disord 151 (2), 639-645.
Wang, P.S., Aguilar-Gaxiola, S., Alonso, J., Angermeyer, M.C., Borges, G., Bromet, E.J., Bruffaerts, R., de Girolamo, G., de Graaf, R., Gureje, O., Haro, J.M., Karam, E.G., Kessler, R.C., Kovess, V., Lane, M.C., Lee, S., Levinson, D., Ono, Y., Petukhova, M., Posada-Villa, J., Seedat, S., Wells, J.E., 2007a. Use of mental health services for anxiety, mood, and substance disorders in 17 countries in the WHO world mental health surveys. Lancet 370 (9590), 841-850.
Wang, P.S., Angermeyer, M., Borges, G., Bruffaerts, R., Chiu, W.T., de Girolamo, G., Fayyad, J., Gureje, O., Haro, J.M., Huang, Y.Q., Kessler, R.C., Kovess, V., Levinson, D., Nakane, Y., Browne, M.A.O., Ormel, J.H., Posada-Villa, J., Aguilar-Gaxiola, S., Alonso, J., Lee, S., Heeringa, S., Pennell, B.E., Chatterji, S., Ustun, T.B., Conso, W.W.M.H.S., 2007b. Delay and failure in treatment seeking after first onset of mental disorders in the World Health Organization's World Mental Health Survey Initiative. World Psychiatry 6 (3), 177-185.

Wang, Y.P., Gorenstein, C., 2013. Psychometric properties of the Beck Depression Inventory-II: a comprehensive review. Braz J Psychiatry 35 (4), 416-431.

Ware, J.E., Jr., Kosinski, M., Bayliss, M.S., McHorney, C.A., Rogers, W.H., Raczek, A., 1995. Comparison of methods for the scoring and statistical analysis of SF-36 health profile and summary measures: summary of results from the Medical Outcomes Study. Med Care 33 (4 Suppl), AS264-279.

Figure legends:

Figure 1. Exploratory factor analyses with oblimin rotation and maximum likelihood estimation (factor loadings of all items) in (a) men and (b) women

Figure 2. Generalized Partial Credit Model information test results.

Figure 3. Generalized Partial Credit Model item trace lines.

Table 1. Sample characteristics

		Men	Women	p
Age (years)	Mean (SD)	54.3 (15.6)	53.7 (15.9)	0.196
Age groups in 2017	18-24	23 (40.4)	34 (59.6)	0.510
	25-34	241 (42.8)	322 (57.2)	
	35-44	295 (42.2)	404 (57.8)	
	45-54	338 (44.1)	428 (55.9)	
	55-64	434 (46.6)	498 (53.4)	
	65-74	415 (46.0)	488 (54.0)	
	75-84	136 (43.5)	177 (56.5)	
	85 -	25 (37.9)	41 (62.1)	
Marital status	single	457 (38.3)	735 (61.7)	<0.001
	married	1450 (46.7)	1657 (53.3)	
Highest education	low	1049 (49.6)	1066 (50.4)	<0.001
	high	858 (39.3)	1326 (60.7)	
Mental health service use (12 months)	no	1812 (45.3)	2184 (54.7)	<0.001
	yes	95 (31.4)	208 (68.6)	
Self-rated health	Mean (SD)	2.1 (0.9)	2.1 (0.9)	0.613
GHQ-12 Item 1	0	1621 (85.0)	1938 (81.1)	<0.01
	1	286 (14.9)	454 (18.9)	
GHQ-12 Item 2	0	1626 (85.2)	1865 (77.9)	<0.01
	1	281 (14.7)	527 (22.1)	
GHQ-12 Item 3	0	1729 (90.6)	2128 (88.9)	0.07
	1	178 (9.3)	264 (11.1)	
GHQ-12 Item 4	0	1749 (91.7)	2131 (89.1)	<0.01
	1	158 (8.2)	261 (10.9)	
GHQ-12 Item 5	0	1416 (74.2)	1548 (64.7)	<0.01
	1	491 (25.7)	844 (35.3)	
GHQ-12 Item 6	0	1700 (89.1)	2048 (85.6)	<0.01
	1	207 (10.8)	344 (14.4)	
GHQ-12 Item 7	0	1662 (87.1)	1974 (82.5)	<0.01
	1	245 (12.8)	418 (17.5)	
GHQ-12 Item 8	0	1749 (91.7)	2110 (88.2)	<0.01
	1	158 (8.2)	282 (11.8)	
GHQ-12 Item 9	0	1643 (86.1)	1901 (79.4)	<0.01
	1	264 (13.8)	491 (20.6)	
GHQ-12 Item 10	0	1742 (91.3)	2099 (87.7)	<0.01
	1	165 (8.6)	293 (12.3)	

Table 1. Continue

		Men	Women	p
GHQ-12 Item 11	0	1756 (92.1)	2138 (89.4)	<0.01
	1	151 (7.9)	254 (10.6)	
GHQ-12 Item 12	0	1715 (89.9)	2067 (86.4)	<0.01
	1	192 (10.1)	325 (13.6)	
MHI-5 Item 1	Mean (SD)	87 (15.6)	85 (17.0)	<0.01
MHI-5 Item 2	Mean (SD)	94 (13.9)	92 (14.9)	<0.01
MHI-5 Item 3	Mean (SD)	72 (20.4)	68 (20.0)	<0.01
MHI-5 Item 4	Mean (SD)	87 (16.5)	82 (17.9)	<0.01
MHI-5 Item 5	Mean (SD)	69 (21.3)	68 (22.0)	0.64
BDI-6 Item 1	0	1641 (86.0)	1887 (78.8)	<0.01
	1	202 (10.5)	388 (16.2)	
	2	63 (3.4)	113 (4.7)	
	3	1 (0.05)	4 (0.1)	
BDI-6 Item 2	0	1749 (91.7)	2106 (88.0)	<0.01
	1	54 (2.8)	73 (3.0)	
	2	89 (4.6)	198 (8.3)	
	3	15 (0.7)	15 (0.7)	
BDI-6 Item 3	0	1581 (82.9)	1916 (80.1)	0.09
	1	49 (2.5)	65 (2.7)	
	2	270 (14.1)	396 (16.5)	
	3	7 (0.3)	15 (0.6)	
BDI-6 Item 4	0	1564 (82.0)	1685 (70.4)	<0.01
	1	273 (14.3)	528 (22.0)	
	2	65 (3.4)	168 (7.0)	
	3	5 (0.2)	11 (0.4)	
BDI-6 Item 5	0	1696 (88.9)	1998 (83.5)	<0.01
	1	184 (9.6)	325 (13.5)	
	2	25 (1.3)	67 (2.8)	
	3	2 (0.1)	2 (0.0)	
BDI-6 Item 6	0	1525 (79.9)	1803 (75.3)	<0.01
	1	327 (17.1)	478 (19.9)	
	2	55 (2.8)	106 (4.4)	
	3	0 (0.0)	5 (0.2)	

Note. p-value denotes the statistical difference between men and women

Appendix 1. Brief 6-item Beck Depression Inventory (items adapted from original 21-item BDI (Beck et al. 1961))

1.
Depressed mood

0	I do not feel sad
1	I feel blue or sad
2a	I am blue or sad all the time and I can't snap out of it
2b	I am so unhappy that it is very painful
3	I am so sad or unhappy that I can't stand it

2.
Pessimism

0	I am not particularly pessimistic or discouraged about the future
1	I feel discouraged about the future
2a	I feel I have nothing to look forward to
2b	I feel that I won't ever get over my troubles
3	I feel that the future is hopeless and that things cannot improve

3.
Dissatisfaction

0	I am not particularly dissatisfied
1	I feel bored most of the time
2a	I feel I have accomplished very little that is worthwhile or that means anything
2b	As I look back on my life all I can see is a lot of failures
3	I feel that I am a complete failure as a person (parent, husband, wife)

4.
Guilt

0	I don't feel particularly guilty
1	I feel bad or unworthy practically a good part of the time
2a	I feel quite guilty
2b	I feel bad or unworthy practically all the time now
3	I feel as though I am very bad or worthless

5.

Self-dislike

0	I don't feel disappointed in myself
1a	I am disappointed in myself
1b	I don't like myself
2	I am disgusted with myself
3	I hate myself

6.

Indecisiveness

0	I make decisions about as well as ever
1	I am less sure of myself now and try to put off making decisions
2	I can't make decisions any more without help
3	I can't make any decisions at all any more