Short running title: Psychometric properties of the online PHQ-4

Psychometric properties of the online version of the Patient Health Questionnaire-4 in a large Colombian sample: Results from the PSY-COVID study

Juan P. Sanabria-Mazo[1, 2, 3, *], Andrés Gómez-Acosta[4], John Alexander Castro-Muñoz[4], Yonatan Ferney Rojas[4], Albert Feliu-Soler[3, **], Juan V. Luciano[1, 2, **], Antoni Sanz [3, *, **]

[1] Institut de Recerca Sant Joan de Déu, Esplugues de Llobregat (España)

[2] Parc Sanitari Sant Joan de Déu, St. Boi de Llobregat (España)

[3] Universitat Autònoma de Barcelona, Cerdanyola del Vallès (España)

[4] Fundación Universitaria Sanitas, Bogotá (Colombia)

***Correspondence**

Juan P. Sanabria-Mazo. Teaching, Research, and & Innovation Unit - Parc Sanitari Sant Joan de Déu. C/ Dr. Antoni Pujadas, 42. 08830. St. Boi de Llobregat, Spain. E-mail: jp.sanabria@pssjd.org

Antoni Sanz. Departament de Psicología Bàsica, Evolutiva i de l'Educació, Universitat Autònoma de Barcelona. Edifici B, Carrer de la Fortuna, Campus de la UAB, 08193, Cerdanyola del Vallès, Spain. E-mail: antonio.sanz@uab.cat

** These authors share senior authorship.

Short running title: Psychometric properties of the online PHQ-4

ABSTRACT

Objectives: The Patient Health Questionnaire-4 (PHQ-4) is a short screening instrument to measure depressive and anxious symptoms. This study evaluated the psychometric properties of the online version of the PHQ-4 in a large sample of the general population in Colombia.

Methods: Data were collected during the first phase of mobility restrictions occasioned by the COVID-19 pandemic. A total of 18'061 adult participants completed the online version of the PHQ-4. The characteristics of the items and subscales were explored. Dimensionality was examined using confirmatory factor analysis (CFA), including an examination of factor invariance across gender and age. Reliability indices were computed and known-groups validity was addressed by estimating associations between PHQ-4 scores and socio-demographic characteristics.

Results: The CFA showed high fit indices for the expected two-factor structure, being invariant across gender and age. Internal consistency was satisfactory for PHQ-2 (α = .83), GAD-2 (α = .79), and PHQ-4 (α = .86). We observed higher scores on depression (PHQ-2), anxiety (GAD-2), and distress (PHQ-4) in women and young people, and those respondents with lower income, unemployed, and lower level of education.

Conclusion: The findings indicate that the PHQ-4 is a reliable and valid screen for depression and anxiety among Colombian people, being recommendable this tool for online surveys.

KEYWORDS

Patient Health Questionnaire, anxiety, depression, COVID-19 pandemic, psychometric properties.

Short running title: Psychometric properties of the online PHQ-4

1. INTRODUCTION

Anxiety disorders and major depression have been identified as the most prevalent emotional disturbances (Wittchen et al., 2011). The high comorbidity between them, close to 50% of cases (Kessler et al., 2015), as well as the impact of both mood disorders on the functioning and quality of life of affected individuals are priority problems for public health systems (Kroenke et al., 2009). The prevalence of anxiety disorders increased to 11% from 1990 to 2010, growing from 200 million to 272 million reported cases worldwide (Baxter et al., 2014). In addition, a meta-analysis that examined 68 studies conducted in 30 countries, between 1994 and 2014, reported a prevalence of depression of around 13% in the general population (Lim et al., 2018). According to World Health Organization (WHO, 2011) population reports estimate that by 2030 emotional disorders will be the most disabling mental health conditions worldwide.

In recent years, it has been highlighted the importance of short, rapid, and reliable screening tools to facilitate diagnosis mood disorders in health care settings (Olariu et al., 2015). Different authors have proposed the use of short screening tools to reduce misdiagnosis (Castro-Rodríguez et al., 2015), optimise health system resources (Cano-Vindel et al., 2018), and improve clinical outcomes (Goldberg et al., 2017). In this context, Patient Health Questionnaire (PHQ-4) is one of the most widely used short screening instruments to measure depressive and anxious symptoms in patients (Kroenke et al., 2009). This self-report instrument combines the two first depression items of the PHQ-9 (Kroenke et al., 2001) and of the Generalized Anxiety Disorder (GAD-7; Spitzer et al., 2006), based on the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) diagnostic criteria for major depressive disorder and generalised anxiety disorder, respectively.

Several studies have demonstrated the goodness-of-fit of the two-factor structure of the PHQ-4 (Kocalevent et al., 2014; Kroenke et al., 2009). Löwe et al. (2010) identified that the PHQ-

4 and PHQ-9 have highly similar psychometric properties. Although the PHQ-4 is a brief instrument, recent meta-analyses have provided evidence of its adequate psychometric properties for the detection of depressive and anxious symptoms, with sensitivity (.76 to .89) and specificity (.76 to .90) indices at a cut-off of 3 (Mitchell et al., 2016) and GAD-2 (Plummer et al., 2016). However, given the characteristics of the PHQ-4, some authors recommend adjusting the selection of the optimal cut-off point according to the target study population (Kroenke et al., 2009).

Online screening tools are reliable for the detection of mood disorders (Muñoz-Navarro et al., 2017a; 2017b). These questionnaires facilitate data collection and help to avoid limitations of data loss in the classic paper-and-pencil format or response bias in face-to-face interviews. Recently, Cano-Vindel et al. (2018) tested the dimensionality, reliability, and validity of a computerised version of the PHQ-4 in a Spanish sample of 1'052 patients from 28 primary care centres. Results indicated adequate internal consistency for depression (α = .86) and anxiety (α = .76), as well as a good sensitivity (.88 and .90, respectively) and specificity (.61, in both cases). Even though the PHQ-4 has been standardised on a representative sample of 1'500 people from the general Colombian population through face-to-face interviews (Kocalevent et al., 2014) – also with adequate properties –, there is no evidence of the psychometric properties of an online version of this instrument in other Spanish-speaking countries apart from Spain.

In addition, as far as we know, the goodness-of-fit of a bifactorial structure for the PHQ-4 has not been tested. Recently, Tibubos et al. (2021) evaluated the internal structure of the PHQ-9 using confirmatory factor analysis (CFA). The bifactor model yielded an excellent fit to the data, being superior to that obtained in the one- and two-factor models. Two types of latent factors are defined in bifactor models: The first is a general factor in which all items are allowed to load; and the second is composed of specific factors in which the items are distributed by their content. In

the case of the PHQ-4 and following the Clark and Watson's tripartite model (1991), the general factor reflects the shared component of anxiety and depression (general distress/negative affect), whereas the specific factors (depression and anxiety after controlling for the general negative affect factor) represent low positive affect (for depression) and hyperarousal (for anxiety). In bifactor models, all latent factors are mutually uncorrelated, and they help to evaluate whether the computation of the subscale scores is justifiable or whether only the total score should be computed and reported.

In this study the objective was three-fold. First, to examine the goodness-of-fit of the one, two-, and bifactor model of the PHQ-4 in a large sample of the Colombian general population. For the first time, the goodness-of-fit for a bifactor model in the PHQ-4 is tested. Thus, we evaluated whether the PHQ-4 could be modelled with a general factor of distress, as measured by 4 items, and 2 specific factors (low positive affect and high arousal), as measured by two item subsets. In line with the evidence reported in previous studies (Cano-Vindel et al., 2018; Kocalevent et al., 2014; Kroenke et al., 2009; Löwe et al., 2010), we speculate that the two-factor correlated model would have better fit to the data than the other models (Hypothesis 1). Second, we tested configurational invariance of our best-fitting model across gender and age, expecting that the dimensions were invariant across gender and age groups (Hypothesis 2). Third, we evaluated the reliability of the anxiety, depression, and distress scales using different reliability indices. We expected that the PHQ-4 subscales would have the capacity to reliably measure anxiety and depression beyond the reliability index examined (Hypothesis 3). Fourth, we explored the relationship between the PHQ-4 scores with demographic characteristics of the sample (gender, age groups, income level, and education level). Based on the results from previous psychometric studies (Cano-Vindel et al., 2018; Kocalevent et al., 2014; Löwe et al., 2010), we expected that

females, older individuals, those with lower incomes, employed, or with lower levels of education would exhibit higher depressive, anxious, and distress symptoms (Hypothesis 4).

2. METHOD

2.1. Study design

Data analyses of the online version of the PHQ-4 were conducted using the database of the PSY-COVID study in Colombia (Sanabria-Mazo & Sanz, 2021). PSY-COVID is a cross-sectional study that aimed to assess the psychosocial impact of the COVID-19 pandemic in 30 countries. Specifically, this article explored data from the general population residing in Colombia during the first phase of the mobility restrictions.

2.2. Participants

In total, 18'833 people completed the online questionnaire in Colombia, of which 772 were excluded from this analysis because they resided in other countries. Finally, the sample consisted of 18'061 participants from all regions of the country. Inclusion criteria were adults (≥ 18 years old) residing in Colombia during the period in which the data were collected (see Table 1).

Insert Table 1 here

2.3. Procedure

Administration of an anonymous online questionnaire generated with Google Forms® was carried out using the snowball method (May 20th to 20 June 20th, 2020). The survey was distributed through social networks, media, and institutional contacts. Data were collected over four consecutive weeks in order to analyse cross-sectionally the cumulative impact of mobility restrictions on depressive and anxious symptoms in the sample. A panel of 30 international experts in clinical and health psychology validated the online questionnaire. This research was approved

by the Ethical Committee on Animal and Human Experimentation of the Autonomous University of Barcelona (CEEAH-5197).

2.4. Measures

The *Patient Health Questionnaire-4* (PHQ-4) was used to measure depressive and anxious symptoms (Löwe et al., 2010). The two items of PHQ-2 correspond to the symptoms of the DSM-IV diagnostic criteria for major depressive disorder and the two items of GAD-2 to the symptoms of generalised anxiety disorder. This ultra-short version contains 4 items with a 4-point Likert response format, where 0 corresponds to "not at all" and 3 to “nearly every day", and questions are asked in the time frame of the last two weeks. The total score of the PHQ-4 (distress) ranges from 0 to 12 and the specific score of its two subscales (PHQ-2 and GAD-2) ranges from 0 to 6. The cut-off points for detecting probable cases of depression (PHQ-2) or anxiety (GAD-2) is 3 or more for each subscale; and for probable cases of distress (PHQ-4) is 6 or more for total scale (Cano-Vindel et al., 2018; Kocalevent et al., 2014; Löwe et al., 2010).

In addition, a socio-demographic information questionnaire was included to collect data about gender, age group, income level, work status, educational level, and region of residence.

2.5. Data analyses

Initially, characteristics of the PHQ-4 were explored, including item means and standard deviations (SD), skewness and kurtosis, corrected item-total correlations, among items in each subscale, and between items of different subscales. Given the brevity of the scales, these correlations were analysed using the Spearman-Brown correction. The dimensionality of the PHQ-4 was examined through CFA, using maximum likelihood as the estimation method. Regarding dimensionality, we tested a 1) one-factor model with the four items loading on one latent factor; 2) two-factor model including two correlated dimensions; and 3) bifactor model with the four items

saturated with a global latent factor of distress plus two uncorrelated specific factors of anxiety and depression. To estimate reliability, both total internal consistency of the scale (PHQ-4) and the subscales (PHQ-2 and GAD-2) were assessed through Cronbach's α, McDonald's ω, and Guttman's λ^2.

In addition to the Chi-square test, the fit indices were analysed according to Schermelleh-Engel et al. (2003): The Tucker–Lewis's index, the comparative fit index and the root mean square error of approximation with 90% confidence intervals. The factorial invariance of the models was tested by gender and age in comparable subsamples, following the configurational procedures proposed by Brown (2014). Furthermore, a known-groups validity approach was used to estimate associations between PHQ-4 scores and socio-demographic characteristics that have been reported in the literature as risk factors for depression and anxiety. For this purpose, univariate group comparisons were performed with PHQ-2, GAD-2, and PHQ-4 scores as dependent variables through t-test and analysis of variance (ANOVA), considering the Bonferroni adjustment for multiple testing. Effect sizes were classified according to Ferguson (2009): recommended minimum effect size representing a "practically" significant effect for social science data (RMPE; $d \geq 0.41$; $\eta^2 \geq 0.04$), moderate ($d \geq 1.15$; $\eta^2 \geq 0.25$), and strong ($d \geq 2.70$; $\eta^2 \geq 0.64$). Statistical analyses were performed with SPSS-26®, AMOS-5, and JASP®.

3. RESULTS

3.1. Item and scale characteristics

Table 2 shows descriptive analyses of the items, subscales (PHQ-2 and GAD-2) and total scale (PHQ-4). Mean (SD) score of PHQ-2 was 2.28 (1.61), GAD-2 was 2.01 (1.67), and PHQ-4 was 4.29 (3.01). Skewness and kurtosis indices (± 1) denote compliance with univariate normality.

Corrected item-total correlations ranged from.62 to.77. Correlation between the two items of PHQ-2 was r = .64 and between the two items of GAD-2 was .68, while correlation of the items with the items of the other subscale ranged from.46 to .68. PHQ-2 and GAD-2 had a correlation of .64, indicating high overlap between subscales. All the above correlations were statistically significant ($p < .01$).

Insert Table 2 here

3.2. Dimensionality

The fit indices for the correlated two-factor model were better than those obtained for the one-factor model [CFI (.99 vs. .94), TLI (.99 vs. .83), NFI (.99 vs. 94), and RMSEA (.04 vs. .23)], which provides strong support for the adequacy of the original model proposed by Kroenke et al. (2009). We tested the bifactor structure, but did not find convergence for this model. Regarding factor loadings of the tested factor models, in the two-factor model ranged between .71 and .92, and those of the one-factor model between .68 and .83. The results slightly differed between the specific factors of anxiety and depression. In line with our Hypothesis 1, these results confirm that the two-factor correlated model have a better fit to the data than the other models.

Insert Figure 1 here

Comparable subsamples were used to test the factorial invariance of the two-factor correlated model by gender (female = 4'305; male = 4'295) and age groups (≤ 32 years = 9'169; > 32 years = 8'892). As shown in Table 3, no structural differences were identified the best-fitting model according to gender and age, with a Δ CFI lower than .01, which confirm Hypothesis 2.

Insert Table 3 here

3.3. Reliability

Reliability of the PHQ-2 (α = .79, ω = .81, and λ^2 = .80), the GAD-2 (α = .83, ω = .83, and λ^2 = .82), and the PHQ-4 (α = .86, ω = .86, and λ^2 = .86) was above .78 on all calculated indicators. The adequate reliability indices for depression, anxiety, and distress confirms Hypothesis 3.

3.4. Known groups validity

As shown in Table 4, statistically significant differences were found in PHQ-2, GAD-2, and PHQ-4 scores according to gender, age groups, income level, work status, and educational level, but not with the RMPE (d < 0.2 and η^2 < 0.12). Females, younger age, unemployed, and those with lower incomes and educational levels reported the higher depression (PHQ-2), anxiety (GAD-2), and distress (PHQ-4) scores. The higher scores for females, and those with lower incomes and educational levels were consistent with Hypothesis 4. However, they were inconsistent with the higher scores expected for older age and employed.

Insert Table 4 here

4. DISCUSSSION

The findings of the current study provide evidence that the online version of PHQ-4 is a reliable and valid self-administered tool for measuring depressive and anxious symptoms in the general population in Spanish. Previous studies have demonstrated the validity of the classic paper-and-pencil format of the PHQ-4 in clinical (Weihs et al., 2018; Kroenke et al., 2009) and non-clinical samples (Kocalevent et al., 2014; Löwe et al., 2010). However, as far as we know, this is the first study to evaluate the psychometric properties of an online version of the PHQ-4 in a large sample of the general population.

Consistent with previous research carried out in the classic paper-and-pencil format and face-to-face interviews (Kocalevent et al., 2014; Kroenke et al., 2009; Löwe et al., 2010), CFA indicates that the two-factor structure of the online PHQ-4 performs better than the one-factor structure, with excellent fit indices on all parameters (CFI = .99, TLI = .99, NFI = .99, RMSEA = .04). The bifactorial model did not converge probably due to the small number of indicators per latent variable. Therefore, these findings demonstrate that the two-factor correlated model is the best fit to the data, confirming Hypothesis 1. The high correlation between the depression (PHQ-2) and anxiety (GAD-2) subscales is similar to those reported in previous studies (Kroenke et al., 2009; Löwe et al., 2010). Comorbidity between these mood disorders, close to 50% of cases (Kessler et al., 2015), theoretically explains the high correlation identified between both subscales of the PHQ-4 (Kocalevent et al., 2014; Kroenke et al., 2009; Löwe et al., 2010). As in previous validations (Kocalevent et al., 2014; Löwe et al., 2010), the two-factor structure of the PHQ-4 was invariant across gender and age in this study supporting Hypothesis 2.

Reliability values were slightly higher than those reported in other psychometric studies (Cano-Vindel et al., 2018; Kocalevent et al., 2014; Löwe et al., 2010), with values close to .83 for depression (PHQ-2), .79 for anxiety (GAD-2), and .86 for distress (PHQ-4), which supports Hypothesis 3. Although with small effect sizes, the findings of this study provide further evidence about gender, age groups, income level, work status, and educational level role as risk factors for depression and anxiety (Cano-Vindel et al., 2018; Kocalevent et al., 2014; Löwe et al., 2010). In line with other validation studies and Hypothesis 4, it was identified that people of female gender, low income, and low education levels reported higher scores on depression, anxiety, and distress (Cano-Vindel et al., 2018; Kocalevent et al., 2014; Löwe et al., 2010). In contrast, it was found that younger people and unemployed reported higher scores than people who were older (Löwe et

al., 2010) and employed (Kocalevent et al., 2014), reflecting the negative effect of the COVID-19 pandemic on the mental health (Hossain et al., 2020; Serafini et al., 2020).

These findings should be interpreted considering the following limitations. First, the analyses were conducted based on a non-representative sample, which impedes the generalisability of the results to the general population of Colombia or other Spanish-speaking languages. Second, due to the cross-sectional design, it was not possible to calculate the test-retest reliability of the instrument. Third, although the convergent and divergent of the PHQ-4 has been demonstrated in previous studies (Kocalevent et al., 2014; Kroenke et al., 2009; Löwe et al., 2010), no other instruments were used to provide further evidence of construct validity. Fourth, diagnostic interviews were not considered as a procedure to verify criterion validity, making it not possible to provide further evidence on specificity and sensitivity for the optimal cut-off point. Fifth, because of the design, we could not examine the responsiveness, the smallest detectable change, or the minimal clinical important difference for scoring the PHQ-4. Sixth, online data collection can have a negative impact on the representation of population groups with internet connection difficulties, lack of knowledge in the use of new technologies, and low literacy

In conclusion, this study provides further evidence on the reliability and validity of short online screening instruments for the detection of mood disorders. It also demonstrates that presentation in its online format does not alter its psychometric properties. The existing results from the PHQ-2 and GAD-2 denote similar psychometric behaviour to the full versions of the PHQ-9 and GAD-2. Although the PHQ-2 and GAD-2 are reliable subscales for rapid screening of depression and anxiety, the use of their full versions is recommended when all DSM-IV diagnostic criteria need to be assessed. In line with the proposal by Löwe et al. (2010), it is suggested to use the total scale as a global screening tool for distress (PHQ-4), and the depression (PHQ-2) and

anxiety (GAD-2) subscales for their discriminated detection. Finally, PHQ-4 is a short screening tool that can help to optimise the time resources of health systems.

CONFLICT OF INTEREST STATEMENT

All authors declare no conflict of interest.

FUNDING INFORMATION

This research was funded by the Agency for Management of University and Research Grants (AGAUR; 2020PANDE00025). JPS-M has a PFIS predoctoral contract from the Institute of Health Carlos III (ISCIII; FI20/00034). JVL has a "Miguel Servet Type II" contract from the ISCIII (CPII19/00003). AF-S acknowledges the funding from the Serra Húnter program (Generalitat de Catalunya; reference number UAB-LE-8015).

REFERENCES

Baxter, A. J., Vos, T., Scott, K. M., Ferrari, A. J., & Whiteford, H. A. (2014). The global burden of anxiety disorders in 2010. *Psychological Medicine*, *44*(11), 2363. https://doi.org/10.1017/S0033291713003243

Brown, T. A. (2014). *Confirmatory factor analysis for applied research*. Guilford Publications.

Cano-Vindel, A., Muñoz-Navarro, R., Medrano, L. A., Ruiz-Rodríguez, P., González-Blanch, C., Gómez-Castillo, M. D., Capafons-Bonet, A., Chacón, F., Santolayay, F., & PsicAP Research Group. (2018). A computerized version of the Patient Health Questionnaire-4 as an ultra-brief screening tool to detect emotional disorders in primary care. *Journal of Affective Disorders*, *234*, 247-255. https://doi.org/10.1016/j.jad.2018.01.030

Castro-Rodríguez, J. I., Olariu, E., Garnier-Lacueva, C., Martín-López, L. M., Pérez-Solà, V., Alonso, J., Forero, C. G., & INSAyD Investigators. (2015). Diagnostic accuracy and adequacy of treatment of depressive and anxiety disorders: A comparison of primary care and specialized care patients. *Journal of Affective Disorders, 172*, 462-471. https://doi.org/10.1016/j.jad.2014.10.020

Clark, L. A., & Watson, D. (1991). Tripartite model of anxiety and depression: Psychometric evidence and taxonomic implications. *Journal of Abnormal Psychology, 100*, 316–336. https://doi.org/10.1037/0021-843X.100.3.316

Ferguson, C. J. (2016). An effect size primer: a guide for clinicians and researchers. *Professional Psychology: Research and Practice, 40(*5), 532–538. https://doi.org/10.1037/a0015808

Goldberg, D. P., Reed, G. M., Robles, R., Minhas, F., Razzaque, B., Fortes, S., de Jesus, J., Pong-Lam, T., García, J. A., Gask, L., Dowell, A. C., Rosendal, M., Mbatia, J. K., & Saxena, S. (2017). Screening for anxiety, depression, and anxious depression in primary care: A field

study for ICD-11 PHC. *Journal of Affective Disorders, 213*, 199-206. https://doi.org/10.1016/j.jad.2017.02.025

Hossain, M. M., Sultana, A., & Purohit, N. (2020). Mental health outcomes of quarantine and isolation for infection prevention: A systematic umbrella review of the global evidence. *SSRN*, 3561265. http://dx.doi.org/10.2139/ssrn.3561265

Kessler, R. C., Sampson, N. A., Berglund, P., Gruber, M. J., Al-Hamzawi, A., Andrade, L., Bunting, B., Demyttenaere, K., Florescu, S., de Girolamo, G., Gureje, O., He, Y., Hu, C., Huang, Y., Karam, E., Kovess-Masfety, V., Lee, S., Levinson, D., Medina Mora, M. E., Moskalewicz, J., Nakamura, Y., Navarro-Mateu, F., Browne, M. A., Piazza, M., Posada-Villa, J., Slade, T., Ten-Have, M., Torres, Y., Vilagut, G., Xavier, M., Zarkov, Z., Shahly, V., & Wilcox, M. A. (2015). Anxious and non-anxious major depressive disorder in the World Health Organization World Mental Health Surveys. *Epidemiology and Psychiatric Sciences*, *24*(3), 210–226. https://doi.org/10.1017/S2045796015000189

Kocalevent, R. D., Finck, C., Jimenez-Leal, W., Sautier, L., & Hinz, A. (2014). Standardization of the Colombian version of the PHQ-4 in the general population. *BMC Psychiatry*, *14*(1), 1-8. https://doi.org/10.1186/1471-244X-14-205

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine, 16*(9), 606-613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Kroenke, K., Spitzer, R. L., Williams, J. B., & Löwe, B. (2009). An ultra-brief screening scale for anxiety and depression: the PHQ–4. *Psychosomatics, 50*(6), 613-621. https://doi.org/10.1176/appi.psy.50.6.613

Lim, G. Y., Tam, W. W., Lu, Y., Ho, C. S., Zhang, M. W., & Ho, R. C. (2018). Prevalence of depression in the community from 30 countries between 1994 and 2014. *Scientific reports, 8(*1), 2861. https://doi.org/10.1038/s41598-018-21243-x

Löwe, B., Wahl, I., Rose, M., Spitzer, C., Glaesmer, H., Wingenfeld, K., Schneider, A., & Brähler, E. (2010). A 4-item measure of depression and anxiety: validation and standardization of the Patient Health Questionnaire-4 (PHQ-4) in the general population. *Journal of Affective Disorders, 122*(1-2), 86-95. https://doi.org/10.1016/j.jad.2009.06.019

Mitchell, A. J., Yadegarfar, M., Gill, J., & Stubbs, B. (2016). Case finding and screening clinical utility of the patient health questionnaire (PHQ-9 and PHQ-2) for depression in primary care: A diagnostic meta-analysis of 40 studies. *British Journal of Psychiatry Open, 2*(2), 127-138. https://doi.org/10.1192/bjpo.bp.115.001685

Muñoz-Navarro, R., Cano-Vindel, A., Medrano, L. A., Schmitz, F., Ruiz-Rodríguez, P., Abellán-Maeso, C., Font-Payeras, M. A., & Hermosilla-Pasamar, A. M. (2017a). Utility of the PHQ-9 to identify major depressive disorder in adult patients in Spanish primary care centers. *BMC Psychiatry, 17*(1), 291. https://doi.org/10.1186/s12888-017-1450-8

Muñoz-Navarro, R., Cano-Vindel, A., Moriana, J. A., Medrano, L. A., Ruíz-Rodríguez, P., Agüero, L., Rodríguez-Enríquez, M., Pizà, M. R., & Ramírez-Manent, I. (2017b). Screening for generalized anxiety disorder in Spanish primary care centers with the GAD-7. *Psychiatry Research, 256*, 312-317. https://doi.org/10.1016/j.psychres.2017.06.023

Olariu, E., Forero, C. G., Castro-Rodriguez, J. I., Rodrigo-Calvo, M. T., Álvarez, P., Martín-López, L. M., Sánchez-Toto, A., Adroher, N. D., Blasco-Cubedo, M. J., Vilagut, G., Fullana, M. A., & Alonso J. (2015). Detection of anxiety disorders in primary care: A meta-analysis

of assisted and unassisted diagnoses. *Depression and Anxiety, 32*(7), 471-484. https://doi.org/10.1002/da.22360

Plummer, F., Manea, L., Trepel, D., & McMillan, D. (2016). Screening for anxiety disorders with the GAD-7 and GAD-2: A systematic review and diagnostic metaanalysis. *General Hospital Psychiatry, 39,* 24-31. https://doi.org/10.1016/j.genhosppsych.2015.11.005

Sanabria-Mazo, J. P., & Sanz, A. (2021). *PSY-COVID database 1 (Colombia).* Recuperado de https://doi.org/10.6084/m9.figshare.13951775.v3.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online, 8*(2), 23-74.

Serafini, G., Parmigiani, B., Amerio, A., Aguglia, A., Sher, L., & Amore, M. (2020). The psychological impact of COVID-19 on the mental health in the general population. QJM: An International *Journal of Medicine, 113*(8), 531-537. https://doi.org/10.1093/qjmed/hcaa201

Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of Internal Medicine, 166*(10), 1092-1097. https://doi.org/10.1001/archinte.166.10.1092

Tibubos, A. N., Otten, D., Zöller, D., Binder, H., Wild, P. S., Fleischer, T., Johar, H., Atasoy, S., Schulze, L., Ladwig, H. K., Schomerus, G., Linkohr, B., Grabe, H. J., Kruse, J., Schmidt, C. O., Münzel, T., König, J., Brähler, E., & Beutel, M. E. (2021). Bidimensional structure and measurement equivalence of the Patient Health Questionnaire-9: Sex-sensitive assessment of depressive symptoms in three representative German cohort studies. *BMC Psychiatry, 21*(1), 1-13. https://doi.org/10.1186/s12888-021-03234-x

Short running title: Psychometric properties of the online PHQ-4

Weihs, K. L., Wiley, J. F., Crespi, C. M., Krull, J. L., & Stanton, A. L. (2018). Predicting future major depression and persistent depressive symptoms: Development of a prognostic screener and PHQ-4 cutoffs in breast cancer patients. *Psycho-oncology*, *27*(2), 420-426. https://doi.org/10.1002/pon.4472

Wittchen, H. U., Jacobi, F., Rehm, J., Wittchen, H. U., Jacobi, F., Rehm, J., Gustavsson, A., Svensson, M., Jonsson, B., Olesen, J., Allgulander, C., Alonso, J., Faravelli, C., Fratiglioni, L., Jennum, P., Lieb, R., Maercker, A., van Os, J., Preisig, M., Salvador-Carulla, L., Simon, R., & Steinhausen, H. C. (2011). The size and burden of mental disorders and other disorders of the brain in Europe 2010. *European Neuropsychopharmacology*, *21*, 655–679. https://doi.org/10.1016/j.euroneuro.2011.07.018

World Health Organization. (2011). *Global burden of mental disorders and the need for a comprehensive, coordinated response from health and social sectors at the country level. Report by the Secretariat.* World Health Organization.

Short running title: Psychometric properties of the online PHQ-4

Table 1

Socio-demographic characteristics of the sample

Variables, *n* (%)	**Sample (*n* = 18'061)**
Gender	
Female	13'491 (75.0)
Male	4'495 (25.0)
Age groups	
18-24 years	4'564 (25.3)
25-34 years	5'426 (30.0)
35-44 years	3'865 (21.4)
45-54 years	2'411 (13.3)
≥ 55 years	1'795 (9.9)
Income level	
Low	5'173 (28.6)
Medium	11'186 (61.9)
High	1'702 (9.4)
Work status	
Employed	10'872 (60.3)
Unemployed	7'162 (39.7)
Education level	
Primary	322 (1.8)
Secondary	1'536 (8.5)
University	16'185 (89.7)
Region	
Amazon	285 (1.6)
Andean	9'347 (51.9)
Caribbean	1'395 (7.8)
Orinoco	557 (3.1)
Pacific	6'420 (35.7)

Note. *n* = frequency, *%* = percentage.

Table 2

Characteristics of the items and subscales of the PHQ-4

Items	*M (95% CI)*	*SD*	Skewness	Kurtosis	Reliability		
					α	ω	λ^2
Depression (PHQ-2)							
1. Little interest or pleasure in doing things	1.24 (1.23-1.25)	.88	.50	-.37			
2. Feeling down, depressed, or hopeless	1.04 (1.03-1.05)	.89	.66	-.21			
PHQ-2 total score	2.28 (2.26-2.30)	1.61	.61	-.02	.79	.81	.80
Anxiety (GAD-2)							
1. Feeling nervous, anxious, or on edge	1.04 (1.03-1.05)	.89	.64	-.23			
2. Not being able to stop or control worrying	0.97 (0.95-0.98)	.92	.72	-.30			
GAD-2 total score	2.01 (1.98-2.03)	1.67	.74	-.10	.83	.83	.82
Distress (PHQ-4)							
PHQ-4 total score	4.29 (4.24-4.33)	3.01	.73	-.03	.86	.86	.86

Table 3

Test for invariance across gender, age groups, and education level using multi-group CFA

	X^2 (df)	ΔP	CMIN/df	CFI	Δ CFI	RMSEA	Δ RMSEA
Gender							
Female	13.772 (2)		6.886	.999		.036	
Male	12.054 (2)		6.027	.999		.033	
Multigroup analysis							
Configural model	12.91 (2)			.999		.025	
Metric model	13.10 (4)	< .01	3.27	.999	.010	.016	0.012
Scalar model	164.19 (8)	< .01	20.52	.991	.006	.047	0.012
Age groups							
≤ 32 years	15.14 (1)		15.14	.999		.039	
> 32 years	13.54 (1)		13.54	.999		.038	
Multigroup analysis							
Configural model	28.68 (2)		14.34	.999		.027	
Metric model	34.33 (4)	< .01	8.58	.999	.002	.020	0.003
Scalar model	1444.32 (8)	< .01	180.54	.958	.003	.035	0.019

Table 4
Association PHQ-4 scores and socio-demographic characteristics

Variables	**PHQ-2 (0-6)**		**GAD-2 (0-6)**		**PHQ-4 (0-12)**	
	Score ***M*** **(S*D*)**	**Difference groups** ***p*** **value (effect size)**	**Score** ***M*** **(S*D*)**	**Difference groups** ***p*** **value (effect size)**	**Score** ***M*** **(S*D*)**	**Difference groups** ***p*** **value (effect size)**
Gender		< .001 (d = 0.15)		< .001 (d = 0.19)		< .001 (d = 0.19)
Female	2.34 (1.60)		2.08 (1.68)		4.42 (3.00)	
Male	2.09 (1.63)		1.77 (1.63)		3.86 (2.99)	
Age groups		< .001 (η^2 = 0.09)		< .001 (η^2 = 0.03)		< .001 (η^2 = 0.07)
1. 18-24 years	2.98 (1.65)	1 > 2 > 3 > 4 > 5	2.43 (1.81)	1 > 2 > 3 > 4 > 5	5.41 (3.14)	1 > 2 > 3 > 4 > 5
2. 25-34 years	2.38 (1.55)		2.04 (1.68)		4.42 (2.96)	
3. 35-44 years	1.99 (1.50)		1.86 (1.57)		3.85 (2.82)	
4. 45-54 years	1.76 (1.45)		1.70 (1.54)		3.46 (2.75)	
5. ≥ 55 years	1.53 (1.37)		1.55 (1.40)		3.08 (2.51)	
Income level		< .001 (η^2 = 0.03)		< .001 (η^2 = 0.01)		< .001 (η^2 = 0.02)
1. Low	2.69 (1.70)	1 > 2 > 3	2.27 (1.80)	1 > 2 > 3	4.95 (3.23)	1 > 2 > 3
2. Medium	2.15 (1.54)		1.92 (1.61)		4.08 (2.88)	
3. High	1.88 (1.54)		1.76 (1.55)		3.63 (2.81)	
Work status		< .001 (d = 0.15)		< .001 (d = 0.08)		< .001 (d = 0.13)
Employed	2.18 (1.57)		1.95 (1.64)		4.13 (2.94)	
Unemployed	2.43 (1.65)		2.09 (1.72)		4.52 (3.09)	
Education level		= .002 (η^2 < 0.01)		= .004 (η^2 < 0.01)		= .001 (η^2 = 0.01)
1. Primary	2.25 (1.74)		2.06 (1.80)		4.32 (3.28)	
2. Secondary	2.42 (1.71)	2 > 3 > 1	2.14 (1.82)	2 > 3 > 1	4.55 (3.26)	2 > 3 > 1
3. University	2.27 (1.60)		1.99 (1.65)		4.26 (2.98)	

Between-group differences were calculated through t-test and ANOVA, considering the Bonferroni adjustment for multiple testing.