Psychiatry Research 256 (2017) 53–58

Contents lists available at ScienceDirect

Psychiatry Research

journal homepage: www.elsevier.com/locate/psychres

General Health Questionnaire-12 validity in Colombia and factorial equivalence between clinical and nonclinical participants

Francisco J. Ruiz[a,*], Diana M. García-Beltrán[a], Juan C. Suárez-Falcón[b]

[a] Facultad de Psicología, Fundación Universitaria Konrad Lorenz, Bogotá, Colombia
[b] Facultad de Psicología, Universidad Nacional de Educación a Distancia, Madrid, Spain

ARTICLE INFO

Keywords:
General Health Questionnaire-12
Screening
Psychological distress
Emotional symptoms
Measurement invariance
Colombian

ABSTRACT

The General Health Questionnaire – 12 (GHQ-12) is a widely used screening self-report for emotional disorders among adults. However, there is little evidence concerning the validity of the GHQ-12 in Colombia and its factorial invariance between nonclinical and clinical samples. Accordingly, the current study aims to explore the GHQ-12 validity in Colombian nonclinical and clinical samples. The GHQ-12 was administered to a total of 1641 participants, including a sample of undergraduates, one of general population, and a clinical sample. The internal consistency of the GHQ-12 across samples was good (overall alpha of .90). The one-factor model showed a good fit to the data and was considered theoretically more coherent than the two-factor model with positive and negative items loading in separate factors. Metric and scalar invariance were observed across nonclinical and clinical samples. The GHQ-12 scores were strongly and positively related to emotional symptoms and experiential avoidance, and negatively related to life satisfaction. According to the receiver operating characteristic (ROC) curves, a threshold score of 11/12 was optimal to identify emotional disorders. In conclusion, the GHQ-12 is a valid screening self-report in Colombia that provides scores that can be compared across clinical and nonclinical participants.

1. Introduction

The General Health Questionnaire (GHQ; Goldberg and Williams, 1988) is one of the most used mental health screening self-reports. It was designed to be used in non-psychiatric medical consultations to detect changes in patients' functioning. The original version of the GHQ consisted of 60 items that are rated on a 4-point Likert-type scale. However, a 12-item version (i.e., GHQ-12) was subsequently developed and has been adopted due to its brevity and good psychometric properties (Goldberg et al., 1997), being one of the most used screening instruments worldwide (Hewitt et al., 2010).

The GHQ-12 consists of 6 positively and 6 negatively worded items, with response options changing depending on the item type (response options for positively worded items: better, same, worse, and much worse than usual; responses options for negatively worded items: absolutely not, same, more, and much more than usual). Several scoring methods has been used with the GHQ-12, with the two most popular ones being the Likert scoring method (0-1-2-3) and the so-called GHQ method (0-0-1-1). In both cases, higher scores reflect greater levels of psychological distress.

Some debate has been raised regarding the factor structure of the GHQ-12. Although some studies have suggested two- and three-factor models (e.g., Graetz, 1991; Shevlin and Adamson, 2005), Hankins (2008) has provided compelling evidence that the GHQ-12 is unidimensional by demonstrating that additional factors are the product of an artifact of the method of analysis. Factorial equivalence of the GHQ-12 across gender has been found in some studies (e.g., Drapeau et al., 2010) but, to our knowledge, it has not been analyzed across clinical and nonclinical samples. This is an important issue because in the absence of factorial equivalence, scores on the GHQ-12 cannot be compared across clinical and nonclinical individuals.

Several Spanish translations of the GHQ-12 have been conducted, which have shown different psychometric properties (e.g., Campos-Arias, 2007; Rocha et al., 2011; Sánchez-López and Dresch, 2008; Villa et al., 2013). Rocha et al. analyzed the psychometric properties of the GHQ-12 in a very large community sample of 29476 Spanish participants. These authors found Cronbach's alphas between .86 and .90 and, like Hankins (2008), they concluded that the GHQ-12 should be used as a unidimensional scale. In Colombia, Campos-Arias (2007) analyzed the psychometric properties of the GHQ-12 in a large community sample of 2496 individuals. This author found a Cronbach's alpha of .78, which is relatively lower when compared with the internal consistency found for

* Correspondence to: Fundación Universitaria Konrad Lorenz, Carrera 9 bis, No 62-43, Bogotá, Cundinamarca, Colombia.
E-mail address: franciscoj.ruizj@konradlorenz.edu.co (F.J. Ruiz).

http://dx.doi.org/10.1016/j.psychres.2017.06.020
Received 24 December 2016; Received in revised form 23 May 2017; Accepted 10 June 2017
Available online 11 June 2017

the original scale. Additionally, the author did not report which Spanish version was used or how it was translated. Villa et al. administered the Spanish version of the GHQ-12 by Rocha et al. with few modifications to a small sample of 85 hospitalized patients with health problems. They found that the GHQ-12 showed a good internal consistency with a Cronbach's alpha of .84, but they also found that Item 11 did not show an acceptable factor loading.

In conclusion, the psychometric properties of the GHQ-12 remain largely unexplored in Colombia, and further studies are necessary to warrant that the GHQ-12 is a valid screening self-report in Colombian samples. In view of the limitations of the previous studies of the GHQ-12 in Colombia, we selected the version by Rocha et al. (2011) to explore the validity of the GHQ-12 in Colombian samples. Secondary aims of this study were to explore the measurement invariance of the GHQ-12 across clinical and nonclinical samples and to provide an empirical cutoff for differentiating individuals with emotional disorders from nonclinical participants. The comprehensibility of the GHQ-12 items was first explored with a sample of undergraduates and experts on emotional disorders who rated their content validity. Afterward, the GHQ-12 was administered to three samples with a total of 1641 participants.

2. Methods

2.1. Participants

2.1.1. Sample 1

This sample consisted of 925 undergraduates (age range 18–63, $M = 21.37$, $SD = 3.83$) from seven universities of Bogotá. Fifty-six percent of the sample was studying Psychology. The other studies included Law, Engineering, Philosophy, Communication, Business, Medicine, and Theology. Sixty-six percent were women. Of the overall sample, 30.9% of participants had received psychological or psychiatric treatment at some time, but only 5.4% were currently in treatment. Also, 2.9% of participants were taking some psychotropic medication.

2.1.2. Sample 2

The sample consisted of 372 participants (62% females) with age ranging between 18 and 89 years ($M = 26.65$, $SD = 9.81$). The relative educational level of the participants was: 49.2% primary studies (i.e., compulsory education) or mid-level study graduates (i.e., high school or vocational training), 33.4% were undergraduates or college graduates, and 16.4% were currently studying or had a postgraduate degree. They responded to an anonymous internet survey distributed through the Internet and social media (i.e., institutional web-pages, Facebook and Twitter institutional profiles, posts at local Facebook profiles, asking people to share with their contacts, etc.). All participants were Colombian. Forty percent reported having received psychological or psychiatric treatment at some time, but only 7.5% were currently in treatment. Also, 4.3% of participants reported consumption of some psychotropic medication.

2.1.3. Sample 3

It consisted of 344 patients (67.7% of them were women), with an age range of 18–67 years ($M = 28.41$, $SD = 11.23$). Most of the participants were being evaluated in the institutional psychological consultation center (91%), in which inexpensive psychological therapy is offered to general population in Bogotá or in additional private consultations in Bogotá (9%). Most of the participants (79.7%) stated that the reason for consultation were emotional symptoms, 9% sexual disorders, and 11.3% other problems (e.g., couple, familiar, lack of social skills, etc.). Only 7.1% of the participants reported that they were consuming some psychotropic medication.

2.2. Instruments

2.2.1. General Health Questionnaire – 12 (Goldberg and Williams, 1988; Spanish version by Rocha et al., 2011)

The GHQ-12 is a 12-item, 4-point Likert-type scale that is frequently used as screening for psychological disorders. Respondents are asked to indicate the degree to which they have recently experienced a range of common symptoms of distress. Validation studies in 15 countries have found areas under the curve between 82 and 85 (Goldberg et al., 1997).

2.2.2. Depression, Anxiety, and Stress Scales – 21 (DASS-21; Antony et al., 1998; Spanish version by Daza et al., 2002)

The DASS-21 is a 21-item, 4-point Likert-type scale (3 = *applied to me very much, or most of the time,* 0 = *did not apply to me at all*) consisting of sentences describing negative emotional states. It contains three subscales (Depression, Anxiety, and Stress) and has shown good internal consistency and convergent and discriminant validity. The DASS-21 has shown good psychometric properties in Colombia (Ruiz et al., 2017). Strong positive correlations were expected between the GHQ-12 and the DASS subscales. Cronbach's alphas ranged from .86 to .92, .80 to .84, and .80 to .88 for Depression, Anxiety and Stress, respectively.

2.2.3. Acceptance and Action Questionnaire – II (AAQ-II; Bond et al., 2011; Spanish translation by Ruiz et al., 2013)

The AAQ-II is a 7-item, 7-point Likert-type scale (7 = *always,* 1 = *never true*) that measures general experiential avoidance or psychological inflexibility. The items reflect unwillingness to experience unwanted emotions and thoughts and the inability to be in the present moment and behave according to value-directed actions when experiencing unwanted psychological events. The Spanish version by Ruiz et al. (2016b) showed good psychometric properties and a one-factor structure in Colombian samples. In this study, Cronbach's alphas of the AAQ-II ranged from .88 (Sample 1) to .93 (Sample 2). Strong negative correlations were expected between the GHQ-12 and the AAQ-II.

2.2.4. Satisfaction with life survey (SWLS; Diener et al., 1985; Spanish version by Atienza et al., 2000)

The SWLS is a 5-item, 7-point Likert-type scale (7 = *strongly agree,* 1 = *strongly disagree*) that measures self-perceived well-being. Examples of items are “I am satisfied with my life” and “In most ways, my life is close to my ideal.” The SWLS has shown good psychometric properties in Colombia (Ruiz et al., submitted). The Cronbach's alpha of the SWLS in this study was .85. Medium to strong negative correlations were expected between the GHQ-12 and the SWLS.

2.3. Procedures

Before administering the GHQ-12, we conducted two initial studies with the aim of exploring the comprehensibility of the items of the Spanish version by the Rocha et al. (2011) in Colombia and its content validity. Firstly, we administered the GHQ-12 and other questionnaires to 64 clinical psychology trainees in order to analyze the comprehensibility of its items. No understanding problem was mentioned in relation to GHQ-12 items. Secondly, the GHQ-12 items were given to 3 experts in emotional disorders who were asked to rate their representativeness, comprehensibility, interpretation, and clarity. Aiken's V was above the usual threshold of .50 for all GHQ-12 items.

In Sample 1, the administration of the questionnaire package was conducted in the participants' classrooms during the beginning of a regular class. Participants in Sample 2 responded to an anonymous internet survey distributed through the Internet and social media. The survey was called “Survey of Emotional Health in Colombia” and was responded on the platform www.typeform.com. After finishing data collection, a general inform was sent to the participants who provided an email address for that purpose. Afterwards, personal scores and

options for receiving inexpensive psychological treatment were provided when requested by the person. Lastly, participants in Sample 3 responded to the questionnaires during one of the clinical assessment interviews at the beginning of treatment in the presence of their therapist.

All participants provided informed consent and were given a questionnaire packet. Participants in all samples responded to the GHQ-12, DASS-21, and AAQ-II. Additionally, participants in Sample 1 also responded to the SWLS. Upon completion of the study, participants were debriefed about the aims of the study and thanked for their participation. No incentives were provided for participation.

2.4. Data analysis

Prior to conducting factor analyses, data from all samples were examined searching for missing values, which were imputed using the matching response pattern of LISREL© (version 8.71, Jöreskog and Sörbom, 1999), which was the software used to conduct the confirmatory factor analyses (CFA). In this imputation method, the value to be substituted for the missing value of a single case is obtained from another case (or cases) having a similar response pattern over the remaining items of the GHQ-12. Only one value was missing.

A robust diagonally weighted least squares (Robust DWLS) estimation method using polychoric correlations was used to conduct the CFA. We computed the Satorra-Bentler chi-square test and the following goodness-of-fit indexes for the one- and two-factor models: (a) the root mean square error of approximation (RMSEA), (b) the comparative fit index (CFI), and (c) the non-normed fit index (NNFI), (d) the expected cross-validation index (ECVI), and (e) the standardized root mean square residual (SRMR). According to Kelloway (1998) and Hu and Bentler (1999), RMSEA values of .10 represent a good fit, and values below .05 represent a very good fit to the data. For the SRMR, values below .08 represent a reasonable fit, and values below .05 indicate a good fit. With respect to the CFI and NNFI, values above .90 indicate well-fitting models, and values above .95 represent a very good fit to the data. The ECVI was computed to compare the goodness of fit of the one-factor model and the two-factor model, with positive and negative items loading on separate factors. Lower ECVI values indicate better fit to the model.

Additional CFA were performed to test for metric and scalar invariance across samples and gender, following Jöreskog (2005), and Millsap and Yun-Tein (2004). In other words, we analyzed whether the item factor loadings and item intercepts are invariant (i.e., equivalents) across samples and between men and women. The analysis of measurement invariance of latent variables or constructs across groups is relevant because it permits to ensure that comparison on the latent variable across groups are valid (i.e., across clinical and nonclinical samples and gender in this study). In the analysis of measurement invariance, the relative fits of three increasingly restrictive models were compared: the multiple-group baseline model, the metric invariance model, and the scalar invariance model. The multiple-group baseline model allowed the twelve unstandardized factor loadings to vary across the samples and in men and women. The metric invariance model, which was nested within the multiple-group baseline model, placed equality constraints (i.e., invariance) on those loadings across groups. Lastly, the scalar invariance model, which was nested within the metric invariance model, was tested by constraining the factor loadings and the item intercepts to be the same across groups. Equality constraints were not placed on estimates of the factor variances because these are known to vary across groups even when the indicators are measuring the same construct in a similar manner (Kline, 2005). For the model comparison, the RMSEA, CFI, and NNFI indexes between nested models were compared. The more constrained model was selected (i.e., second model versus first model, and third model versus second model) if the following criteria suggested by Cheung and Rensvold (2002) and Chen (2007) were met: (a) the difference in RMSEA (ΔRMSEA) was lower than .01; (b) the differences in CFI (ΔCFI) and NNFI (ΔNNFI) were equal to or greater than −.01.

The remaining statistical analyses were performed on SPSS 20©. Alpha coefficients were computed providing 95% confidence intervals (CI) to explore the internal consistency of the GHQ-12 in Samples 1–3 and in the overall sample. Corrected item-total correlations were obtained to identify items that should be removed because of low discrimination item index (i.e., values below .20). Descriptive data were also calculated, and gender differences in GHQ-12 scores were explored by computing independent sample *t*-tests. To examine criterion validity, scores on the GHQ-12 were compared between participants in Sample 1 and 2 (nonclinical participants) and participants in Sample 3 (clinical participants). Pearson correlations between the GHQ-12 and other scales were calculated to assess convergent construct validity.

Lastly, receiver operating characteristic (ROC) curves were computed for the total nonclinical samples (i.e., Samples 1 and 2), excluding those who stated being in psychological/psychiatric treatment and the participants in the clinical sample who stated that emotional symptoms were the reason for consultation. Both GHQ-12 scores in Likert and GHQ scoring (i.e., 0011) were used as the test variables, whereas belonging to clinical and nonclinical sample was the criterion variable.

3. Results

3.1. Descriptive data and psychometric quality of the items

Table 1 shows the original items of the GHQ-12, their translation into Spanish, the descriptive data and corrected item-total correlations for each sample. All items showed good discrimination, with corrected item-total correlations ranging from .43 to .68 in Sample 1, from .44 to .78 in Sample 2, and from .55 to .74 in Sample 3.

Table 2 shows that the alpha coefficient of the GHQ-12 ranged from .88 (Sample 1) to .91 (Samples 2 and 3), with an overall alpha of .90. In Sample 1, there were statistically significant differences across gender in the GHQ-12, with women showing higher scores ($t = -3.55$, $p < .001$). However, no significant differences were found across gender in Sample 2 ($t = -.45$, $p = .65$) and Sample 3 ($t = .69$, $p = .49$).

3.2. Validity evidence based on internal structure

3.2.1. Dimensionality

The goodness-of-fit values of the one-factor model were: S-B$\chi^2(54) = 603.98$, $p < .01$; CFI = .98, NNFI = .98, SRMR = .05, RMSEA = .079, 90% CI [.073, .085]. The CFI and NNFI values indicated a very good fit to the data, the SRMR a good fit, and the RMSEA a good fit according to Kelloway (1998), but acceptable according to Hu and Bentler (1999). Likewise, the upper 90% CI interval of the RMSEA is above the recommendation of .080. Overall, the fit of the one-factor model was acceptable. Fig. 1 depicts the results of the standardized solution of the one-factor model.

Table 3 shows that the goodness-of-fit indexes were better for the two-factor model than for the one-factor model. However, taking into account that the improvement was not large and that correlation between factors was .90, like Hankins (2008), we considered that selecting the one-factor model would be a more parsimonious decision. Further, the better fit of the two-factor model seems to more resemble a method effect than a theoretically based difference.

3.2.2. Measurement invariance

For measurement invariance analyses, we chose the one-factor model which showed an acceptable fit to the data (see above). Table 4 shows the results of the metric and scalar invariance analyses. Parameter invariance was supported at both the metric and scalar levels across samples and gender because changes in RMSEA, CFI, and NNFI

Table 1
Corrected item-total correlations and descriptive data.

Item	Corrected item-total correlation			*M* (*SD*)		
	Sample 1	Sample 2	Sample 3	Sample 1	Sample 2	Sample 3
1. ¿Ha podido concentrarse bien en lo que hacía?	.52	.55	.57	1.23 (.74)	1.25 (.71)	1.60 (.82)
2. ¿Sus preocupaciones le han hecho perder mucho el sueño?	.43	.44	.55	.97 (.95)	.95 (.98)	1.25 (1.02)
3. ¿Ha sentido que está desempeñando un papel útil en la vida?	.46	.61	.60	.92 (.75)	1.15 (.91)	1.35 (.88)
4. ¿Se ha sentido capaz de tomar decisiones?	.50	.63	.62	.86 (.71)	.98 (.75)	1.29 (.88)
5. ¿Se ha notado constantemente agobiado y en tensión?	.59	.65	.66	1.18 (.94)	1.13 (.98)	1.64 (.92)
6. ¿Ha tenido la sensación de que no puede superar sus dificultades?	.67	.76	.74	.85 (.90)	.83 (.95)	1.44 (.98)
7. ¿Ha sido capaz de disfrutar de sus actividades normales de cada día?	.57	.66	.67	1.02 (.75)	1.08 (.70)	1.39 (.86)
8. ¿Ha sido capaz de hacer frente adecuadamente a sus problemas?	.59	.68	.68	1.02 (.74)	1.09 (.75)	1.36 (.87)
9. ¿Se ha sentido poco feliz o deprimido/a?	.68	.78	.70	1.00 (.92)	1.05 (.98)	1.56 (.91)
10. ¿Ha perdido confianza en sí mismo/a?	.67	.78	.67	.73 (.89)	.85 (1.01)	1.40 (1.02)
11. ¿Ha pensado que usted es una persona que no vale para nada?	.57	.70	.61	.39 (.74)	.52 (.89)	.85 (1.02)
12. ¿Se siente razonablemente feliz considerando todas las circunstancias?	.58	.64	.69	.92 (.72)	.99 (.75)	1.40 (.84)

Table 2
Alpha coefficients and descriptive data across samples.

	Sample 1: Undergraduates (N = 925)	Sample 2: General population online (N = 372)	Sample 3: Clinical (N = 344)	Overall Sample (N = 1641)
Alpha	.88	.91	.91	.90
95% CI	[.86, .89]	[.90, .931]	[.90, .92]	[.89, .91]
Mean score (*SD*)	11.08 (6.37)	11.87 (7.47)	16.54 (7.86)	12.41 (7.29)

were lower than .01.

3.3. Validity evidence based on relationships with other variables

The GHQ-12 showed correlations with all the other assessed constructs in theoretically coherent ways (see Table 5). Specifically, the GHQ-12 showed strong positive correlations with emotional symptoms as measured by the DASS-21 and experiential avoidance. Negative correlations were found between the GHQ-12 and life satisfaction.

Means and standard deviations of the GHQ-12 score for each Colombian sample can be seen in Table 2. Participants' mean score in the clinical sample (Sample 3) was higher than that of participants in Sample 1 ($t = -11.54, p < .001$) and Sample 2 ($t = -8.13, p < .001$).

3.4. Criterion validity

Fig. 2 shows the ROC curves for the Likert and GHQ scoring methods. The results indicate that the GHQ-12 performed better than chance at identifying emotional disorders. The area under the curve (AUC) was .80, 95 CI [.77, .83] for the Likert scoring method and .78, 95 CI [.75, .81] for the GHQ method. Table 6 presents the trade-off between sensitivity (correctly identifying individuals with emotional disorders) and specificity (screening out individuals without emotional disorders) according to different GHQ-12 thresholds, both in Likert and GHQ scoring methods. A threshold score of 11/12 was adequate using the Likert method (sensitivity of .82 and specificity of .63), where as a threshold of 2/3 was found reasonable for the GHQ scoring method

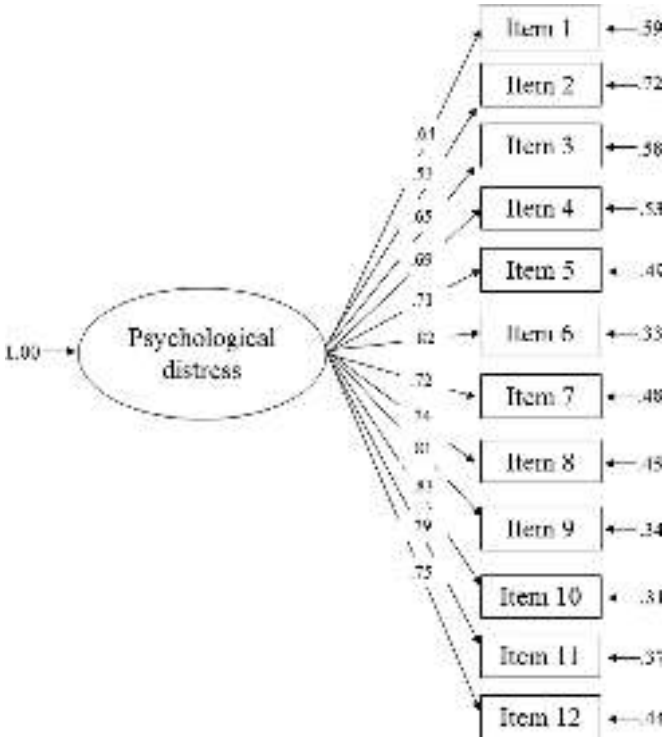

Fig. 1. Standardized solution for the one-factor model conducted with the overall sample.

Table 3
Goodness-of-fit indexes of the one-factor and two-factor model.

Goodness-of-fit indicators	One-factor model	Two-factor model
RMSEA [90% CI]	.079 [.073, .085]	.066 [.060, .071]
CFI	.98	.99
NNFI	.98	.98
ECVI [90% CI]	.40 [.35, .45]	.29 [.25, .33]
SRMR	.05	.04
S-Bχ^2 (*df*)	603.98 (54)	425.71 (53)

Note. CFI = Comparative Fit Index; ECVI = Expected Cross-Validation Index; NFI = Non-Normed Fit Index; RMSEA = Root Mean Square Error of Approximation; S-Bχ^2 = Satorra-Bentler Chi-Square Test; SRMR = Standardized Root Mean Square Residual.

(sensitivity of .80 and specificity of .64).

4. Discussion

The GHQ-12 is one of the most used mental health screening instruments. Several Spanish versions of the GHQ-12 exist, but little evidence has been collected about the validity of the GHQ-12 in Colombia. The current study aimed at advancing in this direction by testing the Spanish version of the GHQ-12 by Rocha et al. (2011) in Colombia. This version was selected because it was validated in an extremely large community sample, showing good psychometric properties. We first confirmed the GHQ-12 items' content validity, according to Colombian experts, and their comprehensibility, according to a sample of undergraduates. Afterward, the GHQ-12 was administered to three samples with a total of 1641 participants (a sample of undergraduates, a sample of online general population, and a clinical sample).

The GHQ-12 showed excellent internal consistency (overall alpha of .90). The one-factor model showed an acceptable fit to the data and was preferred over the two-factor model because the improvement shown by the latter was small and the correlation between factors was .90. Therefore, it seems that the better fit of the two-factor model more resembles a method effect than a theoretically based difference (Hankins, 2008). Measurement invariance at both metric and scalar levels was obtained across samples and gender. This indicates that the GHQ-12 is measuring the same construct across nonclinical and clinical participants, and in men and women.

The GHQ-12 also showed convergent validity in view of the strong positive correlations found with emotional symptoms as measured by the DASS-21 and experiential avoidance, and medium to strong negative correlations with life satisfaction. The GHQ-12 also showed criterion validity to the extent that its scores discriminated between clinical and nonclinical samples. Lastly, the GHQ-12 performed better than chance at identifying emotional disorders. The AUC was slightly lower than in the study by Goldberg et al. (1997) but similar to other subsequent studies (e.g., Baksheev et al., 2011). The threshold scores of 11/12 for the Likert scoring method and 2/3 for the GHQ scoring method were the same as those found by Goldberg et al. (1997) in most of the countries.

Table 5
Pearson correlations between the SWLS scores and other relevant self-report measures.

Measure	S	*r* with GHQ-12
DASS-21 – Depression	1	.66**
	2	.79**
	3	.72**
DASS-21 – Anxiety	1	.47**
	2	.62**
	3	.56**
DASS-21 – Stress	1	.56**
	2	.70**
	3	.67**
AAQ-II	1	.57**
	2	.75**
	3	.64**
SWLS (Life satisfaction)	1	−.44**

*$p < .01$. AAQ-II = Acceptance and Action Questionnaire – II; DASS-21 = Depression, Anxiety and Stress Scale-21; GHQ-12 = General Health Questionnaire – 12; SWLS = Satisfaction with Life Scale.
** $p < .001$.

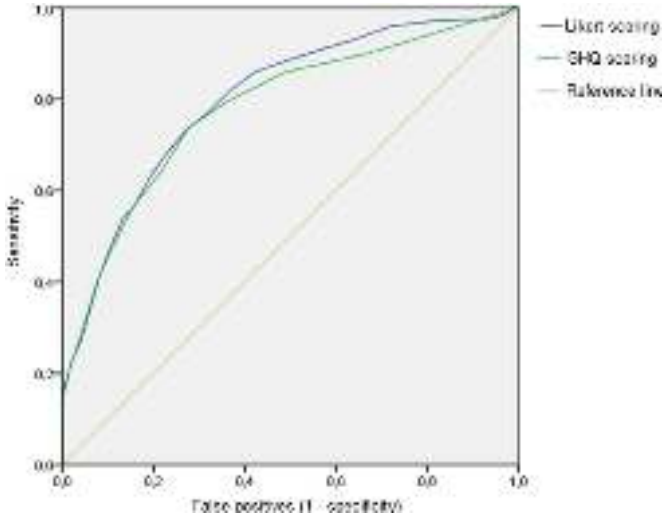

Fig. 2. Receiver operating characteristic (ROC) curve for the GHQ-12 comparing belonging to clinical and nonclinical samples.

One important finding of this study is the factorial equivalence across nonclinical and clinical samples. To our knowledge, this had not been evaluated in previous studies with regard to the GHQ-12. Proof of measurement invariance across nonclinical and clinical samples is important because the studies that use the GHQ-12 usually compare scores from these types of samples. In the absence of data supporting the factorial equivalence of the GHQ-12, the comparison of the scores across these samples is not justified.

The GHQ-12 was administered both via paper-and-pencil (Samples 1 and 3) and the Internet (Sample 2). The results obtained with both

Table 4
Metric and scalar invariance across samples and gender.

Model	S-Bχ^2	df	RMSEA	ΔRMSEA	CFI	ΔCFI	NNFI	ΔNNFI
Measurement invariance across samples								
MG Baseline model	702.77	162	.0782		.976		.971	
Metric invariance	895.41	184	.0842	.006	.969	−.007	.966	−.005
Scalar invariance	1004.85	206	.0843	.000	.965	−.004	.966	.000
Measurement invariance across gender								
MG Baseline model	693.29	108	.0822	–	.979	–	.975	–
Metric invariance	731.31	119	.0801	−.0021	.978	−.001	.976	.001
Scalar invariance	796.51	130	.0799	−.0002	.976	−.002	.976	.000

Table 6
Sensitivity and specificity for selected General Health Questionnaire – 12 (GHQ-12) threshold scores to identify emotional disorders.

GHQ-12 Likert Scoring			GHQ Scoring		
Threshold	Sensitivity	Specificity	Threshold	Sensitivity	Specificity
8/9	91	44	0/1	90	33
9/10	88	51	1/2	86	51
10/11	86	58	2/3	80	64
11/12	82	63	3/4	74	72
12/13	77	68	4/5	64	78
13/14	73	73	5/6	56	85

forms of administration were very similar. This is consistent with previous studies in Spanish samples that did not find significant differences in the administration of the GHQ-28 via the Internet or paper-and-pencil (Vallejo et al., 2007, 2008).

Some limitations of this study are worth mentioning. Firstly, no systematic information was obtained concerning the diagnosis in clinical participants. Secondly, some validity aspects of the GHQ-12 have not been analyzed in the current study (e.g., sensitivity to treatment effects, etc.). However, there is already evidence that the GHQ-12 was sensitive to the treatment effect of a one-session acceptance and commitment therapy protocol focused on reducing repetitive negative thinking (Ruiz et al., 2016a). Thirdly, the percentage of women was significantly higher than the percentage of men in the composition of the samples. This limitation is reduced by the finding of measurement invariance across gender. Fourthly, the percentage of participants in Sample 2 that reported having received psychological or psychiatric treatment seemed high (40%). This could be due to the title given to the survey ("Survey of Emotional Health in Colombia") that could have attracted the attention of people that experienced emotional problems. However, the percentage of undergraduates in Sample 1 that reported having received treatment was also high (30.9%). These high rates could be due to participants failing to differentiate psychological/psychiatric assessment from treatment or the availability of multiple options for receiving inexpensive psychological consultation in Colombia. However, the percentage of participants that were receiving psychological/psychiatric treatment was considerably lower (5.4% and 7.5%, respectively) and typical of nonclinical samples. Lastly, the ROC curves were computed with belonging to nonclinical and clinical sample as the criterion variable. This criterion is not a gold standard, but was the only one available for the current study. Further studies should confirm the threshold scores of the GHQ-12 using diagnostic interviews as criterion variable for computing the ROC curves.

In conclusion, the Spanish version of the GHQ-12 by Rocha et al. (2011) can be used as a screening mental health tool in Colombia. Further studies might explore the psychometric properties of GHQ-12 in other Spanish-speaking countries and test for measurement invariance across countries.

Conflicts of interest

None.

References

Antony, M.M., Bieling, P.J., Cox, B.J., Enns, M.W., Swinson, R.P., 1998. Psychometric properties of the 42-item and 21-item versions of the depression anxiety stress scales (DASS) in clinical groups and a community sample. Psychol. Assess. 10, 176–181. http://dx.doi.org/10.1037/1040-3590.10.2.176.

Atienza, F.L., Pons, D., Balaguer, I., García-Merita, M., 2000. Propiedades psicométricas de la Escala de satisfacción con la vida en adolescentes [Psychometric properties of the satisfaction with life scale in adolescents]. Psicothema 12, 314–319.

Baksheev, G.N., Robinson, J., Cosgrave, E.M., Baker, K., Yung, A.R., 2011. Validity of the 12-item General Health Questionnaire (GHQ-12) in detecting depressive and anxiety disorders among high school students. Psychiatry Res. 187, 291–296.

Bond, F.W., Hayes, S.C., Baer, R.A., Carpenter, K.M., Guenole, N., Orcutt, H.K., et al., 2011. Preliminary psychometric properties of the Acceptance and Action Questionnaire – II: a revised measure of psychological inflexibility and experiential avoidance. Behav. Ther. 42, 676–688. http://dx.doi.org/10.1016/j.beth.2011.03.007.

Campos-Arias, A., 2007. Cuestionario general de Salud-12: Análisis de factores en población general de Bucaramanga, Colombia [General Health Questionnaire-12: analysis of factors in general population of Bucaramanga, Colombia]. Iatreia 20, 29–36.

Chen, F.F., 2007. Sensitivity of goodness of fit indexes to lack of measurement invariance. Struct. Equ. Model. 14, 464–504. http://dx.doi.org/10.1080/10705510701301834.

Cheung, G.W., Rensvold, R.B., 2002. Evaluating goodness-of-fit indexes for testing measurement invariance. Struct. Equ. Model. 9, 233–255. http://dx.doi.org/10.1207/S15328007SEM0902_5.

Daza, P., Novy, D.M., Stanley, M., Averill, P., 2002. The Depression Anxiety Stress Scale-21: Spanish translation and validation with a Hispanic sample. J. Psychopathol. Behav. 24, 195–205.

Diener, E., Emmons, R.A., Larsen, R.J., Griffin, S., 1985. The satisfaction with life scale. J. Pers. Assess. 49, 71–75.

Drapeau, A., Beaulieu-Prévost, D., Marchand, A., Boyer, R., Préville, M., Kairouz, S., 2010. A life-course and time perspective on the construct validity of psychological distress in women and men. Measurement invariance of the K6 across gender. BMC Med. Res. Methodol. 10, 68.

Goldberg, D.P., Gater, R., Sartorius, N., Ustun, T., Piccinelli, M., Gureje, O., et al., 1997. The validity of two versions of the GHQ in the WHO study of mental illness in general health care. Psychol. Med. 27, 191–197.

Goldberg, D., Williams, P., 1988. A User's Guide to the General Health Questionnaire. NFER-Nelson, Windsor, UK.

Graetz, B., 1991. Multidimensional properties of the General Health Questionnaire. Soc. Psychiatry Psychiatr. Epidemiol. 26, 132–138.

Hankins, M., 2008. The reliability of the twelve-item General Health Questionnaire (GHQ-12) under realistic assumptions. BMC Public Health 8, 355.

Hewitt, C.E., Perry, A.E., Adams, B., Gilbod, M., 2010. Screening and case finding for depression in offender populations: a systematic review of diagnostic properties. J. Affect. Disord. 128, 72–82.

Hu, L., Bentler, P.M., 1999. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct. Equ. Model. 6, 1–55. http://dx.doi.org/10.1080/10705519909540118.

Jöreskog, K.G., 2005. Structural Equation Modeling With Ordinal Variables Using LISREL. Scientific Software International, Lincolnwood.

Jöreskog, K.G., Sörbom, D., 1999. LISREL 8.30. Scientific Software International, Chicago.

Kelloway, E.K., 1998. Using LISREL for Structural Equation Modeling: a Researcher's Guide. Sage, Thousand Oaks.

Kline, R.B., 2005. Principles and Practice of Structural Equation Modeling. Guilford Press, New York.

Millsap, R.E., Yun-Tein, J., 2004. Assessing factorial invariance in ordered-categorical measures. Multivar. Behav. Res. 39, 479–515.

Rocha, K., Pérez, K., Rodríguez-Sanz, M., Borrell, C., Obiols, J.E., 2011. Propiedades psicométricas y valores normativos del general health Questionnaire (GHQ-12) en población española [Psychometric properties and normative scores of the General Health Questionnaire (GHQ-12) in general Spanish population]. Int. J. Clin. Health Psychol. 11, 125–139.

Ruiz, F.J., García-Martín, M.B., Suárez-Falcón, J.C., Odriozola-González, P., 2017. The hierarchical factor structure of the Spanish version of the depression anxiety and stress scale-21. Int. J. Psychol. Psychol. Ther. 17, 97–105.

Ruiz, F.J., Langer, A.I., Luciano, C., Cangas, A.J., Beltrán, I., 2013. Measuring experiential avoidance and psychological inflexibility: the Spanish translation of the Acceptance and Action Questionnaire. Psicothema 25, 123–129. http://dx.doi.org/10.7334/psicothema2011.239.

Ruiz, F.J., Riaño-Hernández, D., Suárez-Falcón, J.C., Luciano, C., 2016a. Effect of a one-session ACT protocol in disrupting repetitive negative thinking: a randomized multiple-baseline design. Int. J. Psychol. Psychol. Ther. 16, 213–233.

Ruiz, F.J., Suárez-Falcón, J.C., Cárdenas-Sierra, S., Durán, Y.A., Guerrero, K., Riaño-Hernández, D., 2016b. Psychometric properties of the Acceptance and Action Questionnaire – II in Colombia. Psychol. Rec. 66, 429–437.

Sánchez-López, M.P., Dresch, V., 2008. The 12-item General Health Questionnaire: reliability, external validity and factor structure in the Spanish population. Psicothema 20, 839–843.

Shevlin, M., Adamson, G., 2005. Alternative factor models and factorial invariance of the GHQ-12: a large sample analysis using confirmatory factor analysis. Psychol. Assess. 17, 231–236.

Vallejo, M.A., Jordán, C.M., Díaz, M.I., Comeche, M.I., Ortega, J., 2007. Psychological assessment via the Internet: a reliability and validity study of online (vs paper-and-pencil) versions of the General Health Questionnaire-28 (GHQ-28) and the Symptoms Check-List-90-Revised (SCL-90-R). J. Med. Internet Res. 9 (e2), 1–10.

Vallejo, M.A., Mañanes, G., Comeche, M.I., Díaz, M.I., 2008. Comparison between administration via Internet and paper-and-pencil administration of two clinical instruments: SCL-90-R and GHQ-28. J. Behav. Ther. Exp. Psychiatry 39, 201–208.

Villa, I.C., Zuluaga, C., Restrepo, L.F., 2013. Psychometric properties of the general health Goldberg GHQ-12 Questionnaire applied at a hospital facility in the city of Medellín. Av. Psicol. Latinoam. 31, 532–545.