Psychological Assessment

© 2014 American Psychological Association
1040-3590/14/$12.00 DOI: 10.1037/a0036468

The Impact of Ambiguous Response Categories on the Factor Structure of the GHQ–12

Juan J. Rey and Francisco J. Abad
Universidad Autónoma de Madrid

Juan R. Barrada
Universidad de Zaragoza

Luis E. Garrido
Universidad Iberoamericana

Vicente Ponsoda
Universidad Autónoma de Madrid

Previous research has suggested multiple factor structures for the 12-item General Health Questionnaire (GHQ–12), with contradictory evidence arising across different studies on the validity of these models. In the present research, it was hypothesized that these inconsistent findings were due to the interaction of 3 main methodological factors: ambiguous response categories in the negative items, multiple scoring schemes, and inappropriate estimation methods. Using confirmatory factor analysis with appropriate estimation methods and scores obtained from a large (n = 27,674) representative Spanish sample, we tested this hypothesis by evaluating the fit and predictive validities of 4 GHQ–12 factor models—unidimensional, Hankins' (2008a) response bias model, Andrich and Van Schoubroeck's (1989) 2-factor model, and Graetz's (1991) 3-factor model—across 3 scoring methods: standard, corrected, and Likert. In addition, the impact of method effects on the reliability of the global GHQ–12 scores was also evaluated. The combined results of this study support the view that the GHQ–12 is a unidimensional measure that contains spurious multidimensionality under certain scoring schemes (corrected and Likert) as a result of ambiguous response categories in the negative items. Therefore, it is suggested that the items be scored using the standard method and that only a global score be derived from the instrument.

Keywords: General Health Questionnaire, GHQ–12, confirmatory factor analysis, factor structure

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Mental health problems are highly prevalent in all regions of the world, with 12.2%–48.6% lifetime prevalence rates (World Health Organization [WHO], 2008). It is estimated, for example, that more than 350 million people in the world suffer from depression alone (WHO, 2012). Because early identification and intervention can notably reduce the impact of mental health problems (Härter, Woll, Wunsch, Bengel, & Reuter, 2006), the availability of valid and easy-to-use screening instruments has become a matter of critical importance (Baksheev, Robinson, Cosgrave, Baker, & Yung, 2011; Vodermaier, Linden, & Siu, 2009).

The General Health Questionnaire (GHQ; Goldberg, 1972) is a widely used instrument intended to detect psychological disorders in community and nonpsychiatric clinical settings (Fernandes & Vasconcelos-Raposo, 2013). Although the initial GHQ questionnaire contained 60 Likert-type items, over the years a range of shortened versions have appeared, including the GHQ–30, GHQ–28, GHQ–20, and the GHQ–12 (Goldberg & Williams, 1988). Of these, the 12-item General Health Questionnaire (GHQ–12) has become particularly popular due to its short length, which has made it an attractive assessment tool in large epidemiological studies (Penninkilampi-Kerola, Miettunen, & Ebeling, 2006). The GHQ–12 is composed of six positively phrased items (1, 3, 4, 7, 8, and 12) and six negatively phrased items (2, 5, 6, 9, 10, and 11; see Table 1). The items are answered via a 4-point Likert scale with response categories that vary as a function of the polarity of the items. The categories and scoring values for the *positive* items are 0 = *more than usual*, 1 = *same as usual*, 2 = *less than usual*, and 3 = *much less than usual*, and those for the *negative* items are 0 = *not at all*, 1 = *no more than usual*, 2 = *rather more than usual*, and 3 = *much more than usual*.

Despite the widespread use of the GHQ–12, there is a lot of controversy regarding its factor structure. At the moment, a substantial amount of literature supports one-, two-, and three-factor solutions for the instrument (Aguado et al., 2012; Campbell & Knowles, 2007; Campbell, Walker, & Farrell, 2003). The GHQ–12 was initially designed as a one-factor measure (Goldberg & Williams, 1988), and many authors currently advocate the unidimensional use of the scale (e.g., Gao et al., 2004; Shevlin & Adamson, 2005). However, because the one-factor solution has frequently fitted the empirical data poorly (Gao et al., 2004; Ip & Martin, 2006), several multidimensional models have been proposed and have obtained support through explor-

Juan J. Rey and Francisco J. Abad, Faculty of Psychology, Universidad Autónoma de Madrid; Juan R. Barrada, Faculty of Human and Social Sciences, Universidad de Zaragoza; Luis E. Garrido, Faculty of Psychology, Universidad Iberoamericana; Vicente Ponsoda, Faculty of Psychology, Universidad Autónoma de Madrid.

This research was partially supported by Grants PSI2009-10341 and PSI2012-33343 from the Ministerio de Economia y Competividad, Spain.

Correspondence concerning this article should be addressed to Francisco J. Abad, Facultad de Psicología, Universidad Autónoma de Madrid, c/Iván Pavlov, 6, Madrid 28049, Spain. E-mail: fjose.abad@uam.es

Table 1
Overview of the Most Relevant General Health Questionnaire–12 Factor Models

		Model						
				Andrich & Van Schoubroeck		Graetz		
No.	Item	Unidimensional Factor 1	Hankins[a] Factor 1	Factor 1	Factor 2	Factor 1	Factor 2	Factor 3
1	Have you been able to concentrate on what you were doing?	×	×	×		×		
3	Have you felt that you were playing a useful part in things?	×	×	×		×		
4	Have you felt capable of making decisions about things?	×	×	×		×		
7	Have you been able to enjoy your normal day-to-day activities?	×	×	×		×		
8	Have you been able to face up to your problems?	×	×	×		×		
12	Have you been reasonably happy, all things considered?	×	×	×		×		
2	*Have you lost much sleep over worry?*	×	×		×		×	
5	*Have you felt constantly under strain?*	×	×		×		×	
6	*Have you felt you could not overcome your difficulties?*	×	×		×		×	
9	*Have you been feeling unhappy and depressed?*	×	×		×		×	
10	*Have you been losing confidence in yourself?*	×	×		×			×
11	*Have you been thinking of yourself as a worthless person?*	×	×		×			×

Note. Negative items are shown in italics.
[a] Unifactorial model with correlated errors for the negative items.

atory and confirmatory factor analytic research (e.g., Andrich & Van Schoubroeck, 1989; Doi & Minowa, 2003; Graetz, 1991). An overview of the most relevant factor models for the GHQ–12 is presented in Table 1.

The factor models shown in Table 1 include the theoretical unidimensional model (Goldberg & Williams, 1988) as well as three highly relevant multidimensional models. Hankins (2008a, 2008b) proposed a multidimensional model in which all the items load on a substantive factor, and the errors for the negative items are correlated. According to Hankins (2008a, 2008b), the multidimensionality of the GHQ–12 is the result of method effects related to the response categories of the negative items, and therefore, only one factor should be interpreted substantively. The empirical results have shown that the Hankins model generally obtains levels of fit that are at least as good as those derived from the other multidimensional models that have been proposed (Abubakar & Fischer, 2012; Aguado et al., 2012; Hankins, 2008a, 2008b; Romppel, Braehler, Roth, & Glaesmer, 2013; Smith, Fallowfield, Stark, Velikova, & Jenkins, 2010; Smith, Oluboyede, West, Hewison, & House, 2013; Ye, 2009).

Another multidimensional model that has found substantial support in the GHQ–12 literature is the Andrich and Van Schoubroeck (1989) two-factor model that separates the positive and negative items into distinct substantive dimensions. These factors have been labeled *Social Dysfunction* (positive items) and *General Dysphoria* (negative items; Y. J. Hu, Stewart-Brown, Twigg, & Weich, 2007), and empirical studies have shown that the levels of fit attained with this model are close to those obtained with three-dimensional models (e.g., French & Tait, 2004; Mäkikangas et al., 2006; Shevlin & Adamson, 2005; Ye, 2009). Regarding the three-factor models, the one proposed by Graetz (1991) has received the most support in the literature. The Graetz model builds upon Andrich and Van Schoubroeck's two-factor model by adding a third factor termed *Loss of Confidence* that is composed of negative Items 10 and 11, while the remaining negative items form a second factor that has been labeled as *Anxiety/Depression* (the first factor is the same for both models). The Graetz model, while more complex than the other models, has usually obtained the best levels of fit empirically (Campbell & Knowles, 2007; Campbell et al., 2003; French & Tait, 2004; Gao et al., 2004; Li, Chung, Chui, & Chan, 2009; Mäkikangas et al., 2006; Martin & Newell, 2005; Padrón, Galán, Durbán, Gandarillas, & Rodríguez-Artalejo, 2012; Penninkilampi-Kerola et al., 2006; Salama-Younes, Montazeri, Ismail, & Roncin, 2009; Shevlin & Adamson, 2005).

As can be seen from the previous commentary, one-, two- and three-dimensional models of the GHQ–12 may be supported based on previous factor analytic studies. Although the more complex models have generally produced better levels of fit (Campbell & Knowles, 2007), there is considerable suspicion that the multidimensionality of the GHQ–12 may be the result of spurious effects or methodological factors (Aguado et al., 2012; Hankins, 2008a, 2008b; Wang & Lin, 2011; Ye, 2009). If this were true, the current lack of consensus regarding the factor structure of the GHQ–12 could be a result of the differential impact of these methodological effects, which might be obscuring its underlying factor structure. A review of these sources of bias and their potential interactions is presented next.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Sources of Bias Affecting the GHQ–12 Factor Structure

Ambiguous Response Categories

Hankins (2008a, 2008b) has suggested that the first two response categories of the negative items—*not at all* and *no more than usual*—are ambiguous and generate confusion for the respondents. Following this line, it has been argued that both of these response options apply equally well to respondents wishing to indicate the absence of a negative mood state, a situation that would result in different patterns of association for the negative items (the positive items have different response categories) and would produce spurious multidimensionality.

The dimensional separation between the positive and negative items may also be explained by taking into account the differences in range between their respective response categories. The verbal labels for the response options in the positive items seem to be bipolar (from one pole of the attribute—*more than usual*—to the opposite pole—*much less than usual*), whereas the response options for the negative items seem to be unipolar (from absence of the attribute—*not at all*—to its presence—*much more than usual*). For the negative items, the unipolar format seems to be the logical choice (e.g., *much less than usual* for a negative symptom—like losing confidence—would imply a double negation that could confuse the respondent). Previous research suggests that bipolar response formats may cause problems when one is attempting to measure negative constructs (e.g., dissatisfaction; Davern & Cummins, 2006; Mazaheri & Theuns, 2009). Unfortunately, as being in an unusual top form (for positive items) is not the same as not showing signs of deterioration (for negative items), both types of formats might be contributing to the generation of separate dimensions.

Response bias may also arise due to the *simultaneous* presence of bipolar and unipolar items. In the case of the positive (bipolar) items, the second category—*same as usual*—could easily be interpreted as the zero point (absence of problems), whereas the zero-point for the negative (unipolar) items might be more difficult to establish. For those respondents who interpret the unipolar scale correctly, the absence of problems is indicated by the first category—*not at all*. Alternatively, some respondents may consider the zero-point to be reflected by the second category—*no more than usual*. It is possible, in this latter case, that the zero point of the positive items (second option) might actually *shift* the zero point for the negative items. Furthermore, it should be noted that the verbal labels for the GHQ items go together with numeric labels (0–3) that may be more consistent with this latter interpretation. From the perspective of the respondent, more attention could be directed to the numeric labels, or alternatively, to the verbal labels, thus creating response bias and spurious multidimensionality. The potential difficulties that can arise in the interpretation of the polarity of response scales are well documented (Davern & Cummins, 2006; Mazaheri & Theuns, 2009; Organization for Economic Co-operation and Development, 2013; Russell & Carroll, 1999; Segura & González-Romá, 2003). Additionally, Schwarz (2010) has shown that numerical labels can have a dramatic impact in the psychological interpretation of verbal labels (e.g., in an experiment, *not at all successful* was interpreted to reflect the absence of outstanding achievements when it was associated with a scale with a 0–10 numeric format and to reflect the presence of explicit failures when it was associated with a scale with a −5 to 5 numeric format).

Multiple Scoring Methods

Several scoring methods have been proposed for the GHQ items. With the standard GHQ scoring method (GHQ–0011), the GHQ items are scored dichotomously by collapsing Categories 1 and 2 and scoring them as 0, and collapsing Categories 3 and 4 and scoring them as 1. This is done in order to eliminate the well-known response bias that arises due to some respondents preferring the middle or extreme options regardless of their trait level (Goldberg & Williams, 1988). In addition, Goodchild and Duncan-Jones (1985) proposed a "corrected" scoring method (GHQ–0111), where the 0–0–1–1 method is applied to the positive items, but the negative items are scored 0–1–1–1 by collapsing Categories 2, 3, and 4. In the GHQ–0111 method, an answer of *no more than usual* to a negative item is considered to indicate the presence of a problem rather than good health (Donath, 2001). Some authors have posited that the GHQ–0111 scoring scheme reduces floor effects, provides more normally distributed scale scores, and has better sensitivity and specificity (Donath, 2001; Goodchild & Duncan-Jones, 1985). Finally, in the Likert scoring scheme (GHQ–0123), the response categories are scored incrementally in the typical Likert fashion, or as 0–1–2–3 in this case.

There are several reasons to expect a strong impact of the scoring method in the factor structure of the GHQ–12. On one hand, if there is spurious multidimensionality due to the ambiguity of the categories *not at all* and *no more than usual* in the negative items, it should disappear with the GHQ–0011 scoring method because it collapses these two categories (Hankins, 2008a). On the other hand, the spurious multidimensionality should emerge with the GHQ–0111 and GHQ–0123 scoring methods, because they score these categories differently.

The empirical literature has provided partial support for the scoring method bias hypothesis. First, the unidimensional structure of the GHQ–12 has been supported with GHQ–0011 scoring, with the more complex models providing only equal or marginally better fit (Aguado et al., 2012; Campbell & Knowles, 2007; Campbell et al., 2003) and having extremely high correlations of .80 and .90 between factors (Aguado et al., 2012; Campbell & Knowles, 2007). In addition, with GHQ–0123 scoring, the Graetz and Hankins multidimensional models have produced the best fit (Aguado et al., 2012; Campbell & Knowles, 2007; Hankins, 2008b; Martin & Newell, 2005), while the unidimensional model has been rejected (Aguado et al., 2012; Hankins, 2008b). It should be noted that the differences in fit between the two substantive multidimensional models (Andrich and Van Schoubroeck's and Graetz's) and Hankins' response bias model have been negligible with GHQ–0123 scoring (Aguado et al., 2012; Hankins, 2008b), suggesting that a substantive interpretation of multiple factors is not appropriate.

Despite the previous results, however, the evidence has not been conclusive regarding the scoring method bias hypothesis. For example, with GHQ–0111 scoring, the results have been mixed, with some studies supporting a unidimensional factor structure (Aguado et al., 2012) and others clearly suggesting multidimensionality (Campbell & Knowles, 2007; Campbell et al., 2003;

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Hankins, 2008b; Martin & Newell, 2005). Furthermore, although the differences in fit between the unidimensional and multidimensional models are greatly reduced with GHQ–0011 scoring, the multidimensional models have still produced significantly better levels of fit in some studies (Hankins, 2008b; Martin & Newell, 2005), and the unidimensionality of the instrument has not been supported by the root-mean-square error of approximation (RMSEA) index in some cases (Martin & Newell, 2005).

Inappropriate Estimation Methods

The use of inappropriate statistical methods is another source of bias that has plagued much of the GHQ–12 factor analytical literature (Aguado et al., 2012; Campbell & Knowles, 2007). For example, many of the previous confirmatory factor analysis (CFA) studies have applied maximum-likelihood estimation (e.g., Campbell et al., 2003; Hankins, 2008b; Martin & Newell, 2005), which assumes that the variables are continuous and normally distributed. However, the GHQ–12 has either binary or 4-point observed scores (depending on the scoring method that is used), which is below the minimum of 5 scale points that is recommended in order to treat a variable as continuous (Finney & DiStefano, 2006; Rhemtulla, Brosseau-Liard, & Savalei, 2012). Furthermore, given the target population and item wording, many of the observed item score distributions show very high levels of skewness and kurtosis, making maximum-likelihood estimation clearly unsuitable (Finney & DiStefano, 2006). Thus, the use of inappropriate estimation methods could have led to spurious factors emerging in the GHQ–12 as a result of biased measures of association and model parameters (Aguado et al., 2012). Indeed, in some recent studies, more appropriate estimation methods (e.g., weighted least squares means and variance adjusted, or WLSMV, estimation based on a polychoric correlation matrix) have been applied and have shown that both of the binary scoring methods (GHQ–0011 and GHQ–0111) eliminate the multidimensionality in the GHQ–12 responses (Aguado et al., 2012; Campbell & Knowles, 2007).

In summary, the presence of ambiguous response categories, the application of multiple scoring procedures, and the use of inappropriate estimation methods appears to have contributed to the diverse findings and lack of agreement in the GHQ–12 factor analytic literature. According to a group of studies, the GHQ–12 measures a multifaceted construct, while other studies suggest that the multidimensionality is spurious, and the result of different sources of bias.

Goals of the Current Study

The aim of the present study was to evaluate the dimensionality of the GHQ–12 in a large representative sample through an integrative approach that includes (a) analyzing the interaction between scoring method and model fit, (b) evaluating the impact of method variance on the reliability of the global GHQ–12 scores, and (c) assessing the predictive validity of the scale scores for each factor model and scoring method. We hoped that this integrative approach would help clarify and reconcile the diverse, and often contradictory, findings in the GHQ–12 factor analytic literature.

In order to analyze the interaction between scoring method and model fit, separate CFA were performed for each of the three scoring methods (GHQ–0011, GHQ–0111, and GHQ–0123) and the four-factor models considered (unidimensional, Hankins' response bias model, Andrich and Van Schoubroeck's two-factor model, and Graetz's three-factor model). Thus, a total of 12 CFA models were assessed, allowing for a clearer understanding of the underlying dimensionality of the GHQ–12. Also, as a means to evaluate the impact of method variance on the reliability of the observed GHQ–12 global scores, nonlinear reliabilities were computed for the unidimensional model and for Hankins' response bias model. In the case of Hankins' model, the method variance (in the form of correlated errors between the negative items) was extracted from the common variance of the global scale scores. Finally, the predictive validities of the GHQ–12 scale scores were assessed through simple and multiple linear regressions in order to determine the relative contribution that general and specific factors make in explaining the variance of related constructs. In all cases, the analyses were carried out using appropriate statistical methods for the level of measurement of the input variables.

Method

Participants and Procedure

The data for this study were collected as part of the National Health Survey of Spain (Government of Spain, Ministry of Health, Social Services, and Equality, 2006), which is conducted to evaluate the health of the Spanish population and to determine the impact of the country's public health policies. The sample was composed of 29,478 noninstitutionalized Spaniards who were at least 16 years old and resided throughout the national territory. Trained staff collected the data between June 2006 and June 2007 through interviews performed at the participant's home. In accordance to the sampling design, one adult resident from each dwelling was randomly chosen to participate in the study. The final sample was obtained after reaching 48,382 residences.

The three main reasons for nonresponses in the study sample were (a) absence of the person sampled (44.7%); (b) refusal to participate (27.7%); and (c) an empty dwelling (19.4%; Government of Spain, Ministry of Health, Social Services, and Equality, 2006). Of the total sample with responses to the survey, 690 cases were removed for this study because the answers were not obtained directly from the participant (responses could be provided by an informant from the same home when the person sampled could not respond due to medical reasons). Also, an additional 1,114 cases were removed from the database because the profile contained missing data on the GHQ–12. In consequence, the sample used for the current study was composed of 27,674 subjects (60.9% women and 39.1% men) from ages 16–97 ($M = 50.25$, $SD = 18.31$). In terms of nationality, 93.0% of the participants in the sample held Spanish nationality, 6.3% held non-Spanish nationality, 0.6% held Spanish nationality plus another one, and 0.1% did not provide information regarding nationality. Of those with non-Spanish nationality, 48.0% were from an American country (not including Canada or the United States), 29.7% were from a country member of the European Union, 14.5% from an African country, and 7.8% from other countries.

The protocol for the administration of the GHQ–12 required that the interviewers read aloud each of the GHQ–12 items and collected the verbal responses of the participants. Throughout this process, the participants also had in their hand a copy of the

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

instrument so that they could read along with the interviewer. In addition, the participants were asked to provide only the *numbers* of the categories that reflected their responses to the items. No benefits were given to ensure participation in the survey, and the responses were collected and later distributed, following the confidentiality considerations required by Spanish and European law.

Measures

GHQ–12. The 12-item General Health Questionnaire (GHQ–12) consists of 12 items, each one assessing the severity of a mental problem over the past weeks using a 4-point Likert-type scale. Three scoring methods were considered: standard scoring (GHQ–0011), corrected scoring (GHQ–0111), and Likert scoring (GHQ–0123; see the introduction section for more details on the response formats and scoring methods for the GHQ–12 items). The Spanish version of the questionnaire was administered to the participants in the current sample (Lobo & Muñoz, 1996).

Other survey variables. The National Health Survey of Spain measures several variables related to health, job conditions, social support, and dependence, which according to theory should be related to the GHQ–12 scores. A brief description of these variables follows next:

Presence of chronic disease was computed as the sum of 29 items that evaluated the presence/absence of chronic diseases such as hypertension, malignancies, diabetes, migraine, chronic back pain, and so on. Cronbach's α for the scale scores in the current sample was .75.

Social support was assessed by means of the Duke–UNC Functional Social Support Questionnaire (DUFSS; Broadhead, Gehlbach, de Gruy, & Kaplan, 1988) in its Spanish version (Bellón Saameño, Delgado Sánchez, Luna del Castillo, & Lardelli Claret, 1996a). The DUFSS measures a person's satisfaction with the functional and affective aspects of social support through 11 items that are responded on a scale ranging from 1 = *much less than I would wish* to 5 = *as much as I would wish*. A sample item for this questionnaire is "I have the possibility of talking to someone about my problems at work or at home." Previous research with Spanish samples supports the existence of a general social support factor with adequate internal consistency (Cronbach's $\alpha > .88$) and test–retest reliability ($r > .75$; Bellón et al., 1996a; Cuéllar-Flores & Dresch, 2012). Cronbach's α for the scale scores in the current sample was .90.

Family support was measured with the Apgar Family Questionnaire (AFQ; Smilkstein, 1978) in its adapted Spanish version (Bellón Saameño, Delgado Sánchez, Luna del Castillo, & Lardelli Claret, 1996b). The AFQ measures adult satisfaction with family support in the domains of adaptation, partnership, growth, affection, and resolve, through five items that are responded to on a scale ranging from 0 = *hardly ever* to 2 = *almost always*. A sample item from this questionnaire is "Are you satisfied with the support you receive from your family when you have a problem?" Previous research has supported the one-factor structure of the AFQ with Spanish respondents, with the scale scores showing good internal consistency (Cronbach's alpha = .84) and test–retest reliability ($r > .80$; Bellón Saameño et al., 1996b). Cronbach's α for the scale scores in the current sample was .78.

Functional dependence was computed as the sum of three dichotomous items that evaluated dependence for personal care, housework, and mobility. The items are scored as 0 = nondependent and 1 = dependent. A participant was classified as dependent when he was not autonomous in at least one of the listed activities (eight, thirteen, and six activities were listed for personal care, housework and mobility, respectively). Sample activities were eating (personal care), cleaning the house (mopping the floor, sweeping; housework) and walking for an hour (mobility). This scale was developed specifically for the National Health Survey and was administered only to persons 65 years or older. Cronbach's α for the scale scores in the current sample was .81.

Previous research has found negative correlations between the GHQ scores and social support (e.g., Harrison, Barrow, Gask, & Creed, 1999), and family support (e.g., McNabb, 1983). On the other hand, positive correlations have been found between the GHQ scores and the presence of chronic disease (Verhaak, Heijmans, Peters, & Rijken, 2005) and functional dependency (e.g., Ayuso-Mateos, Lasa, Vázquez-Barquero, Oviedo, & Díez-Manrique, 1999).

Statistical Analysis

CFA was used to compare the fit of the competing factor models (see Table 1): the unidimensional model, Hankins' one-factor model with correlated errors for the negative items, Andrich and Van Schoubroeck's two-factor item wording model, and Graetz's three-factor model. All the models were analyzed using robust weighted least squares means and variance adjusted (WLSMV), based on polychoric (or tetrachoric) correlation matrices estimated from the raw data. The goodness of fit of the models was assessed using the RMSEA index, the Tucker–Lewis index (TLI), and the comparative fit index (CFI). Values of RMSEA $< .06$, TLI $> .95$, and CFI $> .95$ were considered to indicate a good fit to the data (L. Hu & Bentler, 1999; Yu, 2002). All the CFA models were estimated using the Mplus Version 7 software (Muthén, & Muthén, 1998–2012).

The reliability of the global GHQ-12 scores was computed using Green and Yang's nonlinear reliability coefficient (Green & Yang, 2009; Yang & Green, 2010). This reliability coefficient is computed on the basis of polychoric correlations, and it estimates the reliability of the observed sum scores derived from the factorial model. These reliability analyses were carried out for the two-factor models that generate a global GHQ–12 score, the unidimensional model, and Hankins' response-bias model. The nonlinear reliability coefficients were computed using the R Version 2.15.1 software (R Core Team, 2013), with an adaptation of the SAS syntax provided by Green and Yang (2009).

Finally, the predictive validities of the GHQ–12 scale and subscale scores were analyzed using simple and multiple linear regressions on the variables contained in the National Health Survey of Spain that were expected to be related with psychological well-being or distress. For each scoring method, the variance explained by the global score was compared with the variance accounted for by the two or three respective subscales (according to the Andrich and Van Schoubroeck's or Graetz's models). The simple and multiple regression models were computed using SPSS Version 19.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Results

Descriptive Analysis

The frequency distribution for the GHQ–12 items is presented in Table 2. These results show that the pattern of endorsement varies greatly for the positive and negative items. In the case of the positive items, around 84% of the item responses fell in the second category (*same as usual*). This result is not surprising, considering the nonclinical nature of the sample. It appears that healthy people prefer this category instead of the first option (*more than usual*). In contrast, in the case of the negative items, the responses were more equally distributed between the first and second categories (*not at all* and *no more than usual*), with approximately 47% and 37% of the responses falling in these two options, respectively. A comparison of the results obtained for the positive and negative items seems to indicate that the second response category for the negative items (*no more than usual*) also implies a healthy psychological state, a finding that would go against using the corrected GHQ–0111 scoring method. In fact, 90% of the *no more than usual* responses paired with responses to the option *same as usual* of the positive items. Furthermore, the great majority (87%) of the *not at all* responses also paired with responses to the option *same as usual* of the positive items, suggesting that the first two response categories of the negative items reflect the same psychological state.[1]

Table 3 shows the average polychoric correlations between the GHQ–12 items. Polychoric correlations estimate the correlation between the latent variables that are assumed to underlie the observed ordinal variables. For this reason, the polychoric correlations are expected to be the same regardless of the number of response categories, as long as there is no response bias and the sample size is large enough. A look at the results in Table 3 reveals that the polychoric correlations varied considerably between the different scoring methods. For example, the correlations between the positive items were consistently higher for the dichotomous GHQ–0011 and GHQ–0111 scoring methods (.699), in comparison to the GHQ–0123 Likert method (.504). Likewise, the correlations between the negative items were also higher for GHQ–0011 (.714) and GHQ–0111 (.745) in comparison to GHQ–0123 (.622). These results suggest that when four categories are scored, some response options are not truly ordered. In terms of the correlations between the positive and negative items, the GHQ–0011 produced substantially higher correlations (.668) than either the GHQ–0111 (.443) or GHQ–0123 (.417) scoring methods.

Table 2
Endorsement Percentage for the Response Categories of the General Health Questionnaire–12 Items

No.	Item	Response category 0	1	2	3
1	Have you been able to concentrate on what you were doing?	3	81	14	2
3	Have you felt that you were playing a useful part in things?	9	83	7	1
4	Have you felt capable of making decisions about things?	7	86	6	1
7	Have you been able to enjoy your normal day-to-day activities?	5	81	12	2
8	Have you been able to face up to your problems?	5	88	7	1
12	Have you been reasonably happy, all things considered?	8	83	7	1
2	*Have you lost much sleep over worry?*	31	45	21	4
5	*Have you felt constantly under strain?*	32	42	23	4
6	*Have you felt you could not overcome your difficulties?*	41	45	12	2
9	*Have you been feeling unhappy and depressed?*	46	36	16	3
10	*Have you been losing confidence in yourself?*	61	30	8	1
11	*Have you been thinking of yourself as a worthless person?*	72	23	4	1

Note. The negative items are shown in italics. Categories for the positive items: 0 = *more than usual*, 1 = *same as usual*, 2 = *less than usual*, and 3 = *much less than usual*. Categories for the negative items: 0 = *not at all*, 1 = *no more than usual*, 2 = *rather more than usual*, and 3 = *much more than usual*.

Table 3
Average Polychoric (or Tetrachoric) Correlations Between the General Health Questionnaire–12 Items

Correlated items	Scoring method GHQ–0011	GHQ–0111	GHQ–0123
Positive items	.699	.699	.504
Negative items	.714	.745	.622
Positive items with negative items	.668	.443	.417

Note. GHQ–0011 = General Health Questionnaire standard scoring; GHQ–0111 = GHQ corrected scoring; GHQ-0123 = GHQ Likert scoring.

Model Fit

The fit of the four GHQ–12 CFA models across the three scoring methods are shown in Table 4. In general, the results appear to be consistent for all the models regarding the scoring method. First, with Likert GHQ–0123 scoring, none of the models showed an acceptable level of fit, especially with regard to the RMSEA index that was especially high (> .11). Second, when the corrected GHQ–0111 scoring method was used, all the models except the unidimensional showed a good fit to the data. It appears, therefore, that some response bias present with GHQ–0123 scoring was eliminated when the responses were dichotomized according to the corrected method. Third, with the standard GHQ–0011 scoring, all the models, including the unidimensional, obtained an acceptable level of fit. In the case of the multidimensional models, the fit with GHQ–0011 scoring was very similar to the one obtained with the GHQ–0111 scoring method. In contrast, the unidimensional model had a substantial improvement of fit when GHQ–0011 was used (GHQ–0011: RMSEA = .051, CFI = .982, TLI = .978 vs. GHQ–0111: RMSEA = .117, CFI = .933, TLI = .918). It is worth noting that all the multidimensional models separate the positive and negative items, either by placing them on different factors (Andrich and Van Schoubroeck, and Graetz) or by adding correlated errors to the negative items (Hankins). Thus,

[1] These results were obtained through contingency tables, which were not included in the article due to space constraints.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 4
Fit Measures for the General Health Questionnaire–12 Confirmatory Factor Models

Model/scoring method	*df*	Model fit χ^2	RMSEA	CFI	TLI
Unidimensional					
GHQ–0011	54	3, 869.2	.051	.982	.978
GHQ–0111	54	20, 426.9	.117	.933	.918
GHQ–0123	54	41, 714.2	.167	.876	.848
Hankins					
GHQ–0011	39	1, 802.8	.040	.992	.986
GHQ–0111	39	1, 154.4	.032	.996	.994
GHQ–0123	39	14, 031.3	.114	.958	.929
Andrich and Van Schoubroeck					
GHQ–0011	53	3, 345.7	.047	.985	.981
GHQ–0111	53	3, 780.4	.050	.988	.985
GHQ–0123	53	20, 855.1	.119	.938	.923
Graetz					
GHQ–0011	51	2, 935.1	.045	.987	.983
GHQ–0111	51	2, 931.8	.045	.990	.988
GHQ–0123	51	18, 595.9	.115	.945	.928

Note. All the chi-square (χ^2) goodness-of-fit tests were statistically significant at $p < .00$; *df* = degrees of freedom; RMSEA = root-mean-square error of approximation; CFI = comparative fit index; TLI = Tucker–Lewis index; GHQ–0011 = General Health Questionnaire standard scoring; GHQ–0111 = GHQ corrected scoring; GHQ–0123 = GHQ Likert scoring.

some bias might have still been present in the negative items with GHQ–0111 scoring that was eliminated when the standard scoring method GHQ–0011 was used.

The interfactor correlations for the GHQ–12 CFA models are shown in Table 5. These results reveal that the correlations between factors are generally high (from .558 to .943) but also that the scoring method has a strong impact in the strength of the correlations. For example, with GHQ–0011 scoring, the factor correlations were extremely high (from .894 to .943), indicating an almost complete lack of discriminant validity between the factors of the multidimensional models. In contrast, with corrected GHQ–0111 scoring, the interfactor correlations were substantially lower between factors composed of positive and negative items, respectively (from .558 to .623) but not between factors that only contained negative items (Graetz model: $r_{F2F3} = .913$). These results indicate that scoring differently the first two response categories of the negative items reduces the level of convergence between the positive and negative factors. In the case of Likert GHQ–0123 scoring, the factor correlations were somewhere between the two other scoring methods (from .717 to .848).

Global Score Reliability

The nonlinear reliability estimates for the global GHQ–12 scores were uniformly high when the unidimensional model was analyzed (.899, .909, and .916, for GHQ–0011, GHQ–0111, and GHQ–0123 scoring, respectively). In comparison, with the Hankins model (where the method variance is extracted from the common variance), the reliabilities of the global scores were consistently lower (.853, .566, and .672, for GHQ–0011, GHQ–0111 and GHQ-0123 scoring, respectively). However, whereas the reduction in reliability for GHQ–0011 scoring was relatively small (.046), the decrease was substantial for the GHQ–0111 (.343) and GHQ–0123 (.244) scoring methods. These results indicate that there are large method effects for the GHQ–0123 and GHQ–0111 scoring methods, but not for GHQ–0011 scoring.

Predictive Validity

The predictive validities of the global and subscale GHQ–12 scores are shown in Table 6. These results indicate that the predictive validities of the global scores were comparable across the three scoring methods (average $R^2 = .105$, .075, and .092, for GHQ–0011, GHQ–0111, and GHQ–0123 scoring, respectively). Also, the gains in predictive validities after disentangling the global score into subscale scores were similar, and *close to zero*, across the different factor models and scoring methods (e.g., when Graetz's three subscales were analyzed, an average $\Delta R^2 = .003$, .027, and .007 was obtained for GHQ–0011, GHQ–0111, and GHQ–0123 scoring, respectively). In general, the global GHQ–12 score had its highest validities for chronic disease (R^2 = from .097 to .152), and functional dependence (R^2 = from .116 to .127),

Table 5
Interfactor Correlations for the General Health Questionnaire–12 Models

Scoring method/factor	Andrich & Van Schoubroeck model		Graetz model		
	Factor 1	Factor 2	Factor 1	Factor 2	Factor 3
GHQ–0011					
Factor 1	1.000		1.000		
Factor 2	.943	1.000	.926	1.000	
Factor 2			.912	.894	1.000
GHQ–0111					
Factor 1	1.000		1.000		
Factor 2	.607	1.000	.623	1.000	
Factor 3			.558	.913	1.000
GHQ–0123					
Factor 1	1.000		1.000		
Factor 2	.741	1.000	.717	1.000	
Factor 3			.717	.848	1.000

Note. GHQ–0011 = General Health Questionnaire standard scoring; GHQ–0111 = GHQ corrected scoring; GHQ–0123 = GHQ Likert scoring.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 6
Predictive Validities of the General Health Questionnaire–12 Global and Subscale Scores

			Scoring method								
			GHQ–0011 R^2			GHQ–0111 R^2			GHQ–0123 R^2		
Criterion variable	*N*	Mean (*SD*)	(Global score)	(SS1, SS2)	(SS1, SS2, SS3)	(Global score)	(SS1, SS2)	(SS1, SS2, SS3)	(Global score)	(SS1, SS2)	(SS1, SS2, SS3)
Chronic disease	27,661	3.2 (3.1)	.152	.152	.152	.097	.129	.130	.136	.137	.138
Social support	27,073	48.4 (7.3)	.077	.077	.077	.045	.067	.068	.056	.057	.058
Family support	27,299	9.2 (1.6)	.062	.062	.062	.041	.057	.057	.052	.052	.052
Functional dependence	7,075	1.0 (1.2)	.127	.137	.142	.116	.145	.152	.124	.137	.146
Average R^2			.105	.107	.108	.075	.100	.102	.092	.096	.099
Average ΔR^2				.002	.003		.025	.027		.004	.007

Note. *N* = sample size; *SD* = standard deviation; GHQ–0011 = General Health Questionnaire standard scoring; GHQ–0111 = GHQ corrected scoring; GHQ–0123 = GHQ Likert scoring; (Global score) = variance explained by the global General Health Questionnaire–12 score; (SS1,SS2) = variance explained by subscales in the Andrich and Van Schoubroeck model; (SS1, SS2, SS3) = variance explained by subscales in the Graetz model; ΔR^2 = incremental variance explained by the subscales.

while the highest incremental validities of the subscale scores were obtained for these same criterion variables but were generally small in magnitude (average ΔR^2 = from .005 to .035).

Discussion

Previous research has suggested multiple factor structures for the GHQ–12, with contradictory evidence arising across different studies on the validity of these models (Aguado et al., 2012; Campbell & Knowles, 2007; Campbell et al., 2003). In the present research, it was hypothesized that these inconsistent findings were partly due to the interaction of three main sources of bias—ambiguous response categories in the negative items, multiple scoring procedures, and inappropriate estimation methods—and the results from this study appear to support this view.

CFA was used in the current study in order to compare the validity of four GHQ–12 factor models across three scoring methods, while employing appropriate estimation methods for ordinal variables. The CFA analyses were carried out for the unidimensional model and for three multidimensional models: Hankins' (2008a, 2008b) one-factor model with correlated errors for the negative items, Andrich and Van Schoubroeck's (1989) two-factor model that separates the positive and negative items into distinct substantive dimensions, and Graetz's (1991) three-factor model that distinguishes between a substantive factor composed of positive items and two substantive factors composed of negative items. In addition, the three scoring methods evaluated were the standard GHQ scoring method (GHQ–0011; Goldberg & Williams, 1988), where items are scored dichotomously by collapsing Categories 1 and 2 and scoring them as 0, and Categories 3 and 4 and scoring them as 1; the corrected scoring method (GHQ–0111; Goodchild & Duncan-Jones, 1985), where the 0–0–1–1 method is applied to the positive items while the negative items are scored 0–1–1–1 by collapsing Categories 2, 3, and 4; and the Likert method (GHQ–0123), where the response categories are scored incrementally in the typical Likert fashion, or as 0–1–2–3 in this case.

The general picture obtained from this study is that the GHQ–12 is a unidimensional measure that contains spurious multidimensionality under certain scoring schemes as a result of ambiguous response categories in the negative items. The problematic categories are *not at all* and *no more than usual*, and the problem appears to be that "healthy people" (as inferred from the responses to the positive items) use these two categories indistinctively. This is evidenced by the fact that 87% of the *not at all* responses and 90% of the *no more than usual* responses pair with responses to the same option, *same as usual*, of the positive items. Therefore, when the two categories are collapsed, as is done with the standard GHQ–0011 scoring method, the spurious multidimensionality is eliminated. Moreover, when this scoring scheme is used, the interfactor correlations in the multidimensional models become extremely high (0.89–0.94), indicating an almost complete lack of discriminant validity between the factors.

In contrast to the results obtained with GHQ–0011 scoring, when corrected GHQ–0111 or Likert GHQ–0123 scoring is used, the unidimensional model does not fit the data and the correlations between factors that contain positive and negative items are notably reduced (from .558 to .741). Furthermore, none of the models achieve an acceptable level of fit with GHQ–0123 scoring, a result that seems to indicate that some response categories are not functioning properly and cannot be scored incrementally in the typical Likert fashion. Even though the multidimensional models achieve an acceptable level of fit with GHQ–0111 scoring, the response bias appears to still be present as the positive and negative items artificially separate into different factors.

Taken together, the results from this study appear to indicate that the GHQ–0111 and GHQ–0123 scores are contaminated with response bias. This conclusion is supported by the following evidence: (a) the fact that when GHQ–0011 scoring is used, the unidimensional model fits the data well, (b) the finding that the reliability of the global GHQ–12 score is markedly reduced when method effects are considered for the GHQ–0111 and GHQ–0123 scoring methods, but not for GHQ-0011 scoring, and (c) the minimal incremental validities of the subscale scores when predicting external criteria, beyond what the global GHQ–12 score is already able to explain. Overall, these results are congruent with Hankins' (2008a) suggestion that the scores to the negatively worded items of the GHQ–12 contain response bias. In extension, the results from this study indicate that this bias is essentially eliminated when standard GHQ–0011 scoring is used. Therefore, it is recommended that this scoring method be used when scoring

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

the GHQ–12 and that only a global score is derived from the instrument.

References

Abubakar, A., & Fischer, R. (2012). The factor structure of the 12-item General Health Questionnaire in a literate Kenyan population. *Stress and Health, 28,* 248–254. doi:10.1002/smi.1420

Aguado, J., Campbell, A., Ascaso, C., Navarro, P., Garcia-Esteve, L., & Luciano, J. V. (2012). Examining the factor structure and discriminant validity of the 12-Item General Health Questionnaire (GHQ–12) among Spanish postpartum women. *Assessment, 19,* 517–525. doi:10.1177/1073191110388146

Andrich, D., & Van Schoubroeck, L. (1989). The General Health Questionnaire: A psychometric analysis using latent trait theory. *Psychological Medicine, 19,* 469–485. doi:10.1017/S0033291700012502

Ayuso-Mateos, J. L., Lasa, L., Vázquez-Barquero, J. L., Oviedo, A., & Díez Manrique, J. F. (1999). Measuring health status in community surveys: Internal and external validity of the SF–36. *Acta Psychiatrica Scandinavica, 99,* 26–32. doi:10.1111/j.1600-0447.1999.tb05381.x

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 Research, 187,* 291–296. doi:10.1016/j.psychres.2010.10.010

Bellón Saameño, J. A., Delgado Sánchez, A., Luna del Castillo, J. D., & Lardelli Claret, P. (1996a). Validez y fiabilidad del cuestionario de apoyo social funcional Duke–UNC–11 [Validity and reliability of the Duke–UNC–11 Social Support Questionnaire]. *Atención Primaria, 18,* 153–163.

Bellón Saameño, J. A., Delgado Sánchez, A., Luna del Castillo, J. D., & Lardelli Claret, P. (1996b). Validez y fiabilidad del cuestionario de función familiar Apgar-familiar [Validity and reliability of the Family APGAR Questionnaire]. *Atención Primaria, 18,* 289–295.

Broadhead, W. E., Gehlbach, S. H., De Gruy, F. V., & Kaplan, B. H. (1988). The Duke–UNC Functional Social Support Questionnaire. *Medical Care, 26,* 709–723. doi:10.1097/00005650-198807000-00006

Campbell, A., & Knowles, S. (2007). A confirmatory factor analysis of the GHQ–12 using a large Australian Sample. *European Journal of Psychological Assessment, 23,* 2–8. doi:10.1027/1015-5759.23.1.2

Campbell, A., Walker, J., & Farrell, G. (2003). Confirmatory factor analysis of the GHQ–12: Can I see that again? *Australian and New Zealand Journal of Psychiatry, 37,* 475–483. doi:10.1046/j.1440-1614.2003.01208.x

Cuéllar-Flores, I., & Dresch, V. (2012). Validación del cuestionario de Apoyo Social Funcional Duke–UNK–11 en personas cuidadoras [Validation of the Duke–UNC–11 Functional Social Support Questionnaire in caregivers]. *Revista Iberoamericana de Diagnóstico y Evaluación Psicológica, 34,* 89–101.

Davern, M., & Cummins, R. A. (2006). Is life dissatisfaction the opposite of life satisfaction? *Australian Journal of Psychology, 58,* 1–7. doi:10.1080/00049530500125124

Doi, Y., & Minowa, M. (2003). Factor structure of the 12-item General Health Questionnaire in the Japanese general adult population. *Psychiatry and Clinical Neurosciences, 57,* 379–383. doi:10.1046/j.1440-1819.2003.01135.x

Donath, S. (2001). The validity of the 12-item General Health Questionnaire in Australia: A comparison between three scoring methods. *Australian and New Zealand Journal of Psychiatry, 35,* 231–235.

Fernandes, H. M., & Vasconcelos-Raposo, J. (2013). Factorial validity and invariance of the GHQ–12 among clinical and nonclinical samples, *Assessment, 20,* 219–229. doi:10.1177/1073191112465768

Finney, S. J., & DiStefano, C. (2006). Non-normal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 269–314). Greenwich, CT: Information Age.

French, D., & Tait, R. (2004). Measurement invariance in the General Health Questionnaire–12 in young Australian adolescents. *European Child & Adolescent Psychiatry, 13,* 1–7. doi:10.1007/s00787-004-0345-7

Gao, F., Luo, N., Thumboo, J., Fones, C., Li, S., & Cheung, Y. (2004). Does the 12-item General Health Questionnaire contain multiple factors and do we need them? *Health and Quality of Life Outcomes, 2,* 63–68. doi:10.1186/1477-7525-2-63

Goldberg, D. (1972). *The detection of psychiatric illness by questionnaire: A technique for the identification and assessment of non-psychotic psychiatric illness*. New York, NY: Oxford University Press.

Goldberg, D., & Williams, P. (1988). *A user's guide to the General Health Questionnaire*. Slough, United Kingdom: NFER–Nelson.

Goodchild, M., & Duncan-Jones, M. (1985). Chronicity and the General Health Questionnaire. *British Journal of Psychiatry, 146,* 55–61. doi:10.1192/bjp.146.1.55

Government of Spain, Ministry of Health, Social Services, and Equality. (n.d.). *The National Health Survey of Spain 2006*. Retrieved from https://www.msssi.gob.es/estadEstudios/estadisticas/encuestaNacional/encuesta2006.htm

Graetz, B. (1991). Multidimensional properties of the 12-item General Health Questionnaire. *Social Psychiatry and Psychiatric Epidemiology, 26,* 132–138. doi:10.1007/BF00782952

Green, S. B., & Yang, Y. (2009). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika, 74,* 155–167. doi:10.1007/s11336-008-9099-3

Hankins, M. (2008a). The factor structure of the twelve item General Health Questionnaire (GHQ–12): The result of negative phrasing? *Clinical Practice and Epidemiology in Mental Health, 4,* 10–17. doi:10.1186/1745-0179-4-10

Hankins, M. (2008b). The reliability of the twelve-item General Health Questionnaire (GHQ–12) under realistic assumptions. *BMC Public Health, 8,* 355–362. doi:10.1186/1471-2458-8-355

Harrison, J., Barrow, S., Gask, L., & Creed, F. (1999). Social determinants of GHQ score by postal survey. *Journal of Public Health Medicine, 21,* 283–288. doi:10.1093/pubmed/21.3.283

Härter, M., Woll, S., Wunsch, A., Bengel, J., & Reuter, K. (2006). Screening for mental disorders in cancer, cardiovascular and musculoskeletal diseases. Comparison of HADS and GHQ-12. *Social Psychiatry and Psychiatric Epidemiology, 41,* 56–62. doi:10.1007/s00127-005-0992-0

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55. doi:10.1080/10705519909540118

Hu, Y. J., Stewart-Brown, S., Twigg, L., & Weich, S. (2007). Can the 12-item General Health Questionnaire be used to measure positive mental health? *Psychological Medicine, 37,* 1005–1013. doi:10.1017/S0033291707009993

Ip, W. Y., & Martin, C. R. (2006). Psychometric properties of the 12-item General Health Questionnaire (GHQ-12) in Chinese women during pregnancy and in the postnatal period. *Psychology, Health & Medicine, 11,* 60–69. doi:10.1080/13548500500155750

Li, W., Chung, J. O., Chiu, M. M., & Chan, P. S. (2009). The factor structure of the Chinese version of the 12-item General Health Questionnaire in adolescents. *Journal of Clinical Nursing, 18,* 3253–3261. doi:10.1111/j.1365-2702.2009.02905.x

Lobo, A., & Muñoz, P. E. (1996). Versiones en lengua española validadas. In D. Goldberg & P. Williams (Eds.), *Cuestionario de salud general GHQ (General Health Questionnaire): Guía para el usuario de las distintas versions* [General Health Questionnaire: User's guide to the different versions] (pp. 105–115). Barcelona, Spain: Masson.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Mäkikangas, A., Feldt, T., Kinnunen, U., Tolvanen, A., Kinnunen, M.-L., & Pulkkinen, L. (2006). The factor structure and factorial invariance of the 12-item General Health Questionnaire (GHQ-12) across time: Evidence from two community based samples. *Psychological Assessment, 18,* 444–451. doi:10.1037/1040-3590.18.4.444

Martin, C. R., & Newell, R. J. (2005). The factor structure of the 12-item General Health Questionnaire in individuals with facial disfigurement. *Journal of Psychosomatic Research, 59,* 193–199. doi:10.1016/j.jpsychores.2005.02.020

Mazaheri, M., & Theuns, P. (2009). Effects of varying response formats on self-ratings of life-satisfaction. *Social Indicators Research, 90,* 381–395. doi:10.1007/s11205-008-9263-2

McNabb, R. (1983). Family function and depression. *Journal of Family Practice, 16,* 169–170.

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Organization for Economic Co-operation and Development (OECD). (2013). Methodological considerations in the measurement of subjective well-being. In Secretary-General of the OECD (Ed.), *OECD guidelines on measuring subjective well-being* (pp. 61–138). Paris, France: OECD Publishing. doi:10.1787/9789264191655-6-en

Padrón, A., Galán, I., Durbán, M., Gandarillas, A., & Rodríguez-Artalejo, F. (2012). Confirmatory factor analysis of the General Health Questionnaire (GHQ-12) in Spanish adolescents. *Quality of Life Research, 21,* 1291–1298. doi:10.1007/s11136-011-0038-x

Penninkilampi-Kerola, V., Miettunen, J., & Ebeling, H. (2006). A comparative assessment of the factor structures and psychometric properties of the GHQ–12 and the GHQ–20 based on data from a Finnish population-based sample. *Scandinavian Journal of Psychology, 47,* 431–440. doi:10.1111/j.1467-9450.2006.00551.x

R Core Team. (2012). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rhemtulla, M., Brosseau-Liard, P., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17,* 354–373. doi:10.1037/a0029315

Romppel, M., Braehler, E., Roth, M., & Glaesmer, H. (2013). What is the General Health Questionnaire–12 assessing? Dimensionality and psychometric properties of the General Health Questionnaire–12 in a large scale German population sample. *Comprehensive Psychiatry, 54,* 406–413. doi:10.1016/j.comppsych.2012.10.010

Russell, J. A., & Carroll, J. M. (1999). On the bipolarity of positive and negative affect. *Psychological Bulletin, 125,* 3–30. doi:10.1037/0033-2909.125.1.3

Salama-Younes, M., Montazeri, A., Ismail, A., & Roncin, C. (2009). Factor structure and internal consistency of the 12-item General Health Questionnaire (GHQ–12) and the Subjective Vitality Scale (VS), and the relationship between them: A study from France. *Health and Quality of Life Outcomes, 7,* 22–27. doi:10.1186/1477-7525-7-22

Schwarz, N. (2010). Measurement as cooperative communication: What research participants learn from questionnaires. In G. Walford, E. Tucker, & M. Viswanathan (Eds.), *The SAGE handbook of measurement* (pp. 43–60). Thousand Oaks, CA: Sage. doi:10.4135/9781446268230.n4

Segura, S. L., & González-Romá, V. (2003). How do respondents construe ambiguous response formats of affect items? *Journal of Personality and Social Psychology, 85,* 956–968. doi:10.1037/0022-3514.85.5.956

Shevlin, M., & Adamson, G. (2005). Alternative factor models and factorial invariance of the GHQ–12: A large sample analysis using confirmatory factor analysis. *Psychological Assessment, 17,* 231–236. doi:10.1037/1040-3590.17.2.231

Smilkstein, G. (1978). The Family APGAR: A proposal for a family function test and its use by physicians. *Journal of Family Practice, 6,* 1231–1239.

Smith, A. B., Fallowfield, L. J., Stark, D. P., Velikova, G., & Jenkins, V. (2010). A Rasch and confirmatory factor analysis of the General Health Questionnaire (GHQ)–12. *Health and Quality of Life Outcomes, 8,* 45–54. doi:10.1186/1477-7525-8-45

Smith, A. B., Oluboyede, Y., West, R., Hewison, J., & House, A. O. (2013). The factor structure of the GHQ–12: The interaction between item phrasing, variance, and levels of distress. *Quality of Life Research, 22,* 145–152. doi:10.1007/s11136-012-0133-7

Verhaak, P. F. M., Heijmans, M. J. W. M., Peters, L., & Rijken, M. (2005). Chronic disease and mental disorder. *Social Science & Medicine, 60,* 789–797. doi:10.1016/j.socscimed.2004.06.012

Vodermaier, A., Linden, W., & Siu, C. (2009). Screening for emotional distress in cancer patients: A systematic review of assessment instruments. *Journal of National Cancer Institute, 101,* 1464–1488. doi:10.1093/jnci/djp336

Wang, L., & Lin, W. (2011). Wording effects and the dimensionality of the General Health Questionnaire (GHQ–12). *Personality and Individual Differences, 50,* 1056–1061. doi:10.1016/j.paid.2011.01.024

World Health Organization. (2008). *Scaling up for mental, neurological, and substance abuse disorder*. Geneva, Switzerland: World Health Organization, Mental Health Gap Action Program.

World Health Organization. (2012). *Depression is a common illness, and people suffering from depression need support and treatment. WHO marks 20th anniversary of World Mental Health Day* [Note for the media]. Retrieved from http://www.who.int/mediacentre/news/notes/2012/mental_health_day_20121009/en/

Yang, Y., & Green, S. B. (2010). A note on structural equation modeling estimates of reliability. *Structural Equation Modeling, 17,* 66–81. doi:10.1080/10705510903438963

Ye, S. (2009). Factor structure of the General Health Questionnaire (GHQ–12): The role of wording effects. *Personality and Individual Differences, 46,* 197–201. doi:10.1016/j.paid.2008.09.027

Yu, C. Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes*. Los Angeles: University of California.

Received July 16, 2013
Revision received January 31, 2014
Accepted February 27, 2014 ■

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.