BRIEF REPORT

Wording Effects and the Factor Structure of the 12-Item General Health Questionnaire (GHQ-12)

J. Gabriel Molina and Maria F. Rodrigo
Universitat de València

Josep-Maria Losilla and Jaume Vives
Universitat Autònoma de Barcelona

The 12-item version of the General Health Questionnaire (GHQ-12) has become a popular screening instrument with which to measure general psychological health in different settings. Previous studies into the factorial structure of the GHQ-12 have mainly supported multifactor solutions, and only a few recent works have shown that the GHQ-12 was best represented by a single substantive factor when method effects associated with negatively worded items were considered. Confirmatory factor analysis was applied to compare competing measurement models from previous research, including correlated traits, correlated methods approaches, and correlated uniquenesses approaches, to obtain further evidence about the factorial structure of the GHQ-12. This goal was achieved with data from 3,050 participants who completed the GHQ-12 included in the Catalonian Survey of Working Conditions (Catalonian Labor Relations and Quality of Work Department, 2012). The results showed additional evidence that the GHQ-12 has a unidimensional structure after controlling for method effects associated with negatively worded items. Furthermore, we found evidence for our hypothesis about the spurious nature of the 3-factor solution in Graetz’s (1991) model after comparing its fit with that found for alternative models resulting from different combinations of the negatively worded items. An implication of our results is that future research about the factor structure of the GHQ-12 should take method effects associated with negative wording into account in order to avoid reaching inaccurate conclusions about its dimensionality.

Keywords: psychological health, General Health Questionnaire (GHQ-12), method effects, wording effects, confirmatory factor analysis

Supplemental materials: http://dx.doi.org/10.1037/a0036472.supp

The General Health Questionnaire (GHQ) was developed by Goldberg (1972), and it has been widely used as a screening instrument for measuring general psychological health (GPH) in community and nonpsychiatric clinical settings (Goldberg & Williams, 1988). The questionnaire initially included 60 items, but shorter versions with 30, 28, 20, and 12 items have been developed. The shortest version, with 12 items (GHQ-12), is the most popular because of its brevity and ease of administration. Previous studies have reported good reliability and validity of the tests scores of this 12-item version of the GHQ in different samples and countries (Politi, Piccinelli, & Wilkinson, 1994; Rocha, Pérez, Rodríguez-Sanz, Borrell, & Obiols, 2011; Tait, French, & Hulse, 2003). The GHQ-12 has also been included as part of major national surveys, such as the British Household Panel Survey, the Health Survey for England, the Spanish Health Survey, the National Survey of Occupational Stress in Australian Universities, and the Israel Health and Nutrition Survey.

One of the most controversial aspects to be found in the literature about the GHQ-12 concerns the factor structure underlying the responses to this instrument. Although the GHQ-12 was originally developed as a unidimensional scale, this one-factor latent structure has found empirical support in only a few studies (e.g., Banks et al., 1980; Winefield, Coldney, Winefield, & Tiggermann, 1989). Some alternative multidimensional models, mainly with two or three factors, have been proposed as more appropriate. In this sense, the one with the most empirical support is the three-factor model proposed by Graetz (1991) (Campbell & Knowles, 2007; French & Tait, 2004; Gao et al., 2004; Mäkkikangas et al. 2006; Padrón, Galán, Durán, & Gandarillas, 2012;
Penninkilampi-Kerola, Miettunen, & Ebeling, 2006; Shevlin & Adamson, 2005). The three factors in Graetz’s model are named GPH-1 (anhedonia and social dysfunction; 6 items), GPH-2 (anxiety and depression; 4 items), and GPH-3 (loss of confidence; 2 items). It is important to note that the first factor comprises the six positively worded (PW) items, whereas the six negatively worded (NW) items form the other two factors. The bidimensional model, where the six NW and the six PW items in the GHQ-12 are grouped into two factors, has also obtained wide support, especially in studies based on exploratory factor analysis (e.g., Andrich & Van Schouwbroeck, 1989; Gao et al., 2012; Hankins, 2008; Politii et al., 1994; Schmitz et al., 1999). However, the validity and utility of these multifactor measurement models, mainly Graetz’s model, have been questioned (Campbell & Knowles, 2007; French & Taii, 2004; Gao et al., 2004; Shevlin & Adamson, 2005). The most habitual argument against them is that the unidimensional solution has been repeatedly found with high correlations between the factors. For example, the correlations between these three factors ranged from .83 to .90 in Gao et al. (2004), from .76 to .89 in Campbell and Knowles (2007), and from .72 to .84 in Padrón et al. (2012). Another argument used against Graetz’s model has been the low discriminant validity of the factor scores derived from this model (Gao et al., 2004).

A more recent line of research has questioned the multidimensional nature of the GHQ-12. Hankins (2008) argued that multi-factor models are just an artifact from the inclusion of PW and NW items in the questionnaire, so that the controversy about the factorial structure of the GHQ-12 might relate to underlying method effects. Including both types of items has been commonly recommended in textbooks about test design (e.g., Spector, 1992) as a way of reducing a number of response biases, such as acquiescence, disacquiescence and midpoint response styles. The psychometric literature has shown, though, that this mixture of items can create a spurious factorial differentiation whereby the PW items load on one factor and the NW items load on another (Schmitt & Stults, 1985).

Method effects associated with NW items have received special attention in the psychometric literature (e.g., Rosenberg Self-Esteem Scale; Lindwall et al., 2012; Marsh, 1996; Tomás & Oliver, 1999). In the case of the GHQ-12, a few recent works have focused on analyzing wording effects. Hankins’s (2008) pioneering work on a representative English sample found that, after modeling wording effects for the NW items, the unidimensional model fitted better than both the two-factor model (NW vs. PW items) and Graetz’s three-factor model. Working on a sample of Spanish postpartum women, Aguado et al. (2012) found a slightly better fit for the unidimensional model including wording effects than for Graetz’s model, and they proved that the factor scores derived from Graetz’s model provided little effective discrimination between diagnostic groups. Similarly, working with a sample of Chinese university students, Ye (2009) found a good fit for the three models compared (i.e., the two multidimensional models considered previously and a unidimensional model with an additional method factor associated with the NW items); however, an analysis of the discriminant validity of the three models provided greater support for the unidimensional model with a method factor. Abubakar and Fischer (2012) worked with two samples of Kenyan adolescents and adults and found that the unidimensional model that partialed out the effects of negative wording provided the best structure representation of the GHQ-12. Finally, Smith, Oluboyede, West, Hewison, and House (2013) conducted a study with a representative sample of English individuals age 50 and over, and they also concluded that the unidimensional model including wording effects fitted the data better than the unidimensional model, the two-factor model and the Graetz three-factor model. Taken together, these studies have shown that the unidimensional model including method effects definitely has a better fit than the unidimensional model and a fit slightly better or equal to the multidimensional models, suggesting that the latter might just be an artifact due to wording effects.

Two procedures have been widely used in the literature to statistically control method biases: the correlated traits, correlated methods (CTCM) and the correlated traits, correlated uniquenesses (CTCU) confirmatory factor analysis (CFA) models. The advantages derived from the parameterization of method effects in the CTCM model led Lance, Noble, and Scullen (2002) to recommend this model over the CTCU model as long as there are no problems of nonconvergent or inadmissible solutions. Both procedures have been used to deal with method effects in the GHQ-12: Ye (2009) applied the CTCM model, whereas Hanksins (2008); Aguado et al. (2012); and Smith et al. (2013) used the CTCU model. As far as we know, only the study by Abubakar and Fischer (2012) applied both models to analyze the factor structure of the GHQ-12; however, they did not go deeper into the pros and cons of using both procedures.

To clarify all the above review, the first row of Table 1 shows the five CFA models that have been mainly considered in the study of the GHQ-12 factorial structure, and the rest of the table summarizes the results from all the studies where, among the models compared, either the CTCU or the CTCM models (Models 4 and 5, respectively) were considered. The first column in Table 1 provides the reference for these studies and some clues to their sampling design, whereas the other columns report the goodness-of-fit indices obtained for the five models. Empty cells stand for models that were not considered in the corresponding studies.

We hypothesized, in light of the aforementioned evidence, that the unidimensional model including method effects associated with the NW items would have a better fit than the multidimensional models traditionally considered in the literature to explain the factor structure of the GHQ-12. Thus, our main aim in this study was to examine the factor structure of the GHQ-12 by comparing the competing CFA models from previous research, including both the CTCU and the CTCM approaches, with a representative and comprehensive Spanish sample of workers. Moreover, we hypothesized that the good fit obtained by Graetz’s model, the multifactor model with the biggest support in previous studies, might just be an artifact due to an overparameterization associated with multifactor solutions.

Method

Participants

Data from the Second Catalan Survey of Working Conditions (Catalonian Labor Relations and Quality of Work Department, 2012) were used in this study. The survey was designed to yield a representative sample of all employees living in Catalonia (Spain) according to the Eurostat definition of employee (Euro-
Table 1
Fit indexes for the alternative models of the 12–item General Health Questionnaire

<table>
<thead>
<tr>
<th>Measure</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
<th>Model 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>CFI</td>
<td>.98</td>
<td>.97</td>
<td>.96</td>
<td>.97</td>
<td>.97</td>
</tr>
<tr>
<td>RMSEA</td>
<td>.07</td>
<td>.07</td>
<td>.06</td>
<td>.06</td>
<td>.06</td>
</tr>
<tr>
<td>TLI</td>
<td>.96</td>
<td>.96</td>
<td>.96</td>
<td>.96</td>
<td>.96</td>
</tr>
<tr>
<td>SRMR</td>
<td>.05</td>
<td>.05</td>
<td>.05</td>
<td>.05</td>
<td>.05</td>
</tr>
</tbody>
</table>

Note. Competing models tested for the 12–Item General Health Questionnaire in first row. Underlined numbers identify negatively worded items. GPH: General Psychological Health factor; GPH +: General Psychological Health factor for positive items; GPH –: General Psychological Health factor for negative items; GPH–1: Social dysfunction; GPH–2: Anxiety and depression; GPH–3: Loss of confidence; Method–NW: Method factor associated with negatively worded items; df = degrees of freedom; CFI = Comparative Fit Index; RMSEA = Root Mean Square Error of Approximation; CI = Confidence Interval; TLI = Tucker–Lewis Index; SRMR = Standardized Root Mean Square Residuals.

The sample was collected through a random procedure, with municipalities, households, and individuals as sample units in each of the three stages of the sampling design. Data were collected by professional interviewers by means of a computer-assisted personal interviewing (CAPI) technique in private households. The sampling error was 1.63% with a response rate of 95.5%. The sample comprised a total of 3,601 participants (55.4% were women and 44.6% women) with a mean age of 40.5 years (SD = 11.2, range from 17 to 82). Participants were predominantly European (88.4%); 7.8% were American, 2.7% were African, 0.6% were Australian, and 0.5% were Asian. All the analyses of this study were conducted with the response data from the 3,050 participants who responded to the GHQ-12 included in the survey. No significant differences were found between the entire survey sample and the study sample in terms of sex, age, level of studies completed, workplace size, and activity sector; therefore, missing data were assumed to be at random.

Measures

As part of the Second Catalanion Survey of Working Conditions (Catalonian Labor Relations and Quality of Work Department, 2012), respondents completed the GHQ-12, a self-report scale that contains six PW items (e.g., “Have you been able to face up to problems?”) and six NW items (e.g., “Have you been losing confidence in yourself?”). The GHQ-12 was validated in Spain by Lobo and Muñoz (1996). A peculiar characteristic of the GHQ-12 has to do with its differentiated response scale for the PW items (i.e., more than usual; same as usual; less than usual; and much less than usual) and the NW items (i.e., not at all; no more than usual; rather more than usual; and much more than usual). The 4-point scoring scheme (i.e., 0, 1, 2, 3) was applied in our study given that, on the one hand, some empirical studies (Banks et al., 1980; Campbell & Knowles, 2007) have supported this graduated scoring method over the originally proposed dichotomous scoring procedure and, on the other hand, it has been the most widely applied scoring scheme. Thus, total scores in the GHQ-12 ranged from 0 to a maximum of 36, with higher scores indicating lower levels of GPH.

Statistical Analysis

We estimated a series of confirmatory factor models with LISREL 8.70 (Jöreskog & Sörbom, 2004) using the weighted least squares estimator. The first row of Table 1 shows the specification of all of these CFA models. We estimated three models that do not incorporate method effects: the intended GHQ-12 unidimensional measurement model (Model 1); the two-factor model with PW and NW items defining, respectively, each factor (Model 2); and the Graetz three-factor model (Model 3). Two models were estimated to examine the method effect associated with the NW items: Model 4, a unidimensional model with correlated errors (i.e., a CTCU model), and Model 5, a unidimensional model with an additional factor for the NW items (i.e., a CTCM model). The
goodness-of-fit indices that had been most commonly applied in previous works were computed: the comparative fit index (CFI); the Tucker–Lewis index (TLI); the root mean square error of approximation (RMSEA) with its 90% confidence interval; and the standardized root mean square residual (SRMR). Values greater than .05 for the CFI and TLI and lower than .06 and .08 for the RMSEA and SRMR, respectively, were considered to indicate good model fit.

Results

The frequency distributions for the PW items were quite similar, as were those for the NW items; however, a clear difference appeared when the groups of items were compared. For every PW item, the response category with the highest frequency was Same as usual, whereas for all the NW items it was Not at all. Accordingly, the means were higher for the PW items (average $M = 2.99$) than for the NW items (average $M = 2.40$). On the contrary, the variability in the responses was higher for the NW items (average $SD = 0.62$) than for the PW items (average $SD = 0.55$). As regards the bivariate relationships between items, there were important differences between the PW and the NW items in terms of the value of the polychoric correlations. Thus, the pairwise correlations between the PW items ranged from 0.25 to 0.61 (average correlation $= 0.43$), whereas those between the NW items ranged from 0.40 to 0.85 (average correlation $= 0.61$).

The goodness-of-fit statistics obtained in our work for the five models compared are shown in the last row of Table 1. Comparing the fit of the models without method effects (Models 1, 2, and 3), one can observe that Model 1 showed the worst fit (i.e., the four goodness-of-fit indices did not reach the cutoff criteria considered), as it did in most of the studies compared; only Aguado et al. (2012) reported a good fit for this one-factor model. Although Models 2 and 3 both showed a reasonable fit, it was marginally better for Model 3, as reported in previous research. None of the three models, though, had an SRMR value below the usual cutoff value of .08. There was a high degree of correlation between the three factors in Model 3 (for a description of the factors, see our discussion of Graetz’s model in the introduction): $r_{GPH-1,GPH-2} = .66$; $r_{GPH-1,GPH-3} = .55$; $r_{GPH-2,GPH-3} = .86$. Similar results, obtained in previous studies (see the introduction), have suggested that these factors are not independent and that a more parsimonious solution might be attained. An additional aspect to be noted is that, in most of these studies, the highest relationship corresponds to the two factors containing the NW items (GPH-2 and GPH-3). This raised the issue of whether the improved fit in Model 3 compared to Model 2 might come from just considering an additional factor in the model, taking into account the general rule that the addition of any factor to a model can improve the overall fit of these 10 models. If so, the alternative model (the Graetz model) could be explained by the artificial grouping of NW items. Moreover, the criticism commonly aimed at the Graetz model (see the introduction) has been reinforced here with an additional argument. Thus, it was shown how alternative structurally equivalent models, where the PW items were kept fixed and the NW items were randomly grouped into two factors, fitted the data as well as the Graetz model, which provides further support for the lack of substantive meaning for the two factors contained in the GW items. We conclude, in light of the results of the present study, that the good fit found for multidimensional models in the literature is due to the artificial grouping of NW versus PW items. An immediate implication of this conclusion is that future research about the factor structure of the GHQ-12 should take method effects associated with negative wording into account.
account in order to avoid reaching inaccurate conclusions about its dimensionality. Another implication of this result has to do with a likely bias in the estimation of the relationships between the GHQ scores and a number of covariates (see Podskoﬀ, MacKenzie, & Podskoﬀ, 2012, for a further review of the effects that method biases have on individual measures and on the covariation between different constructs).

As far as the discussion about the CTCU and the CTCM parameterizations of method effects is concerned, it is worth noting that we did not ﬁnd any kind of discussion about the pros and cons of using both frameworks in the context of the GHQ-12. If we delve deeper into the reasons for the better ﬁt of the CTCU model (Model 4) than the CTCM model (Model 5) in our study, a possible explanation comes from the fact that method effects are not explicitly modeled in CTCU models, so different method effects may underlie the observed correlated uniquenesses; on the contrary, any hypothesized method effect will be modeled as a separate latent variable in CTCM models. As a consequence, CTCM models will give rise to better speciﬁed and more parsimonious solutions; yet, if relevant method effects are missing in the model, this will necessarily result in a solution with a worse ﬁt than that derived from a CTCU model. Thus, a further inspection of the relationships between the GHQ-12 items and, more speciﬁcally, between the adjacent equally worded items showed that the correlations between adjacent items were much higher for the pairs of adjacent equally worded items (average correlation = .71) than for the pairs of adjacent items worded in opposite directions (average correlation = .27). These results suggest the presence of an additional method effect consisting of a kind of response inertia between adjacent equally worded items that deserves speciﬁc research in the future.

A relevant question has to do with the nature of the wording effects in the GHQ-12. This issue is even more challenging for this questionnaire than for others, given the different response scale used for the PW and the NW items in the GHQ-12. As hypothesized by Hankins (2008), the response bias could result from the different response scales used for the NW items and the PW items more than from other explanatory factors. Further research should address the speciﬁc contribution of both aspects (i.e., NW items and a different response scale) to the presence of method effects. It is not possible to distinguish either aspect in the usual correlational research using the GHQ-12, and, consequently, (quasi)experimental research would be required to address this issue.

An important strength of this work derived from the quality criteria associated with the survey design of the Catalan Survey of Working Conditions (i.e., sampling framework, random sampling, face-to-face administration by professional interviewers at home, high response rate). As a consequence, this study was based on a representative and comprehensive sample, in contrast to most previous studies about method effects in the responses to GHQ-12, which were based on limited target populations and/or nonprobability sampling. The above characteristics of our sample can be considered relevant: on the one hand, representativeness supports the generalization of the factorial structure found in the sample to our target population; on the other hand, the fact of relying on a comprehensive sample can be positively considered if we take into account that the GHQ-12 is widely used as a screening instrument in the general population. However, given the use of the GHQ-12 in not only community but also clinical settings, further research is needed to test if its factorial structure remains invariant in a clinical sample. Moreover, research about the measurement invariance across age, sex, country, ethnicity, or educational level would be welcomed in order to support comparisons of GPH scores between those groups.

A limitation of this study is that we focused on modeling method effects associated with NW items but not with PW items. The studies on wording effects in different personality and social psychology scales have traditionally paid much more attention to the presence of negative wording. For the widely studied Rosenberg Self-Esteem Scale, the majority of studies have found that method effects are primarily associated with NW items (DiStefano & Motl, 2006; Horan, DiStefano, & Motl, 2003; Marsh, 1996; Tomás & Oliver, 1999). Some recent studies, though, have found a better ﬁt for models including method effects from both NW and PW items (Marsh, Scalas, & Nagengast, 2010; Quilty, Oakman, & Risko, 2006; Wu, 2008). In the context of the GHQ-12, more studies are needed to better understand the method effects associated with positive and negative wording and to quantify the relative importance of different potential sources of method effects (e.g., positive wording, negative wording, and response inertia effects).

References
Gao, W., Stark, D., Bennett, M. I., Siegert, R. J., Murray, S., & Higginson, I. J. (2012). Using the 12-item General Health Questionnaire to screen psychological distress from survivorship to end-of-life care: Dimension-


