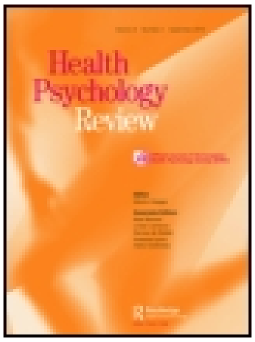


Routledge
Taylor & Francis Group

Health Psychology Review

ISSN: 1743-7199 (Print) 1743-7202 (Online) Journal homepage: http://www.tandfonline.com/loi/rhpr20

The Structure of the General Health Questionnaire (GHQ-12): Two Meta-Analytic Factor Analyses

Timo Gnambs & Thomas Staufenbiel

To cite this article: Timo Gnambs & Thomas Staufenbiel (2018): The Structure of the General Health Questionnaire (GHQ-12): Two Meta-Analytic Factor Analyses, Health Psychology Review, DOI: 10.1080/17437199.2018.1426484

To link to this article: https://doi.org/10.1080/17437199.2018.1426484

Accepted author version posted online: 11 Jan 2018.

Publisher: Taylor & Francis & Informa UK Limited, trading as Taylor & Francis Group

Journal: *Health Psychology Review*

DOI: 10.1080/17437199.2018.1426484

The Structure of the General Health Questionnaire (GHQ-12): Two Meta-Analytic Factor Analyses

Timo Gnambs

Leibniz Institute for Educational Trajectories

Thomas Staufenbiel

Osnabrück University

Author Note

Timo Gnambs, Leibniz Institute for Educational Trajectories, Wilhelmsplatz 3, 96047 Bamberg, Germany, Phone: +49 (0)951 863-3420, Email: timo.gnambs@lifbi.de. Thomas Staufenbiel, Institute of Psychology, Osnabrück University, Seminarstrasse 20, 49074

Osnabrück, Germany, Phone: +49 (0)541 969-4512, Email: thomas.staufenbiel@uni-osnabrueck.de.

Correspondence concerning this article should be addressed to Timo Gnambs.

Abstract

The General Health Questionnaire (GHQ-12) is a popular measure of psychological distress. Despite its widespread use, an ongoing controversy pertains to its internal structure. Although the GHQ-12 was originally constructed to capture a unitary construct, empirical studies identified different factor structures. Therefore, this study examined the dimensionality of the GHQ-12 in two independent meta-analyses. The first meta-analysis used summary data published in 38 primary studies (total N = 76,473). Meta-analytic exploratory factor analyses identified two factors formed by negatively and positively worded items. The second meta-analysis included individual responses of 410,640 participants from 84 independent samples. Meta-analytic confirmatory factor analyses corroborated the two-dimensional structure of the GHQ-12. However, bifactor modeling showed that most of the variance was explained by a general factor. Therefore, subscale scores reflected rather limited unique variance. Overall, the two meta-analyses demonstrated that the GHQ-12 is essentially unidimensional. It is not recommended to use and interpret subscale scores because they primarily reflect general mental health rather than distinct constructs.

Keywords: mental health, distress, factor analysis, meta-analysis, wording effects

The Structure of the General Health Questionnaire (GHQ-12): Two Meta-Analytic Factor Analyses

The General Health Questionnaire (GHQ) is a self-report measure of psychological distress (Goldberg, 1972) that is extensively administered in epidemiological surveys as well as other community and clinical settings (see Fryers et al., 2004). Particularly, its short form with 12 items (GHQ-12) exhibits considerable appeal as a quick and unobtrusive screening instrument to identify people with minor psychological disturbance being at risk of developing psychiatric disorders (Goldberg & Williams, 1988). Despite its popularity, the structure of the instrument is still subject to an ongoing debate. Originally, the GHQ-12 was assumed to capture a single trait. Although some empirical studies supported this assumption (e.g., Fernandes & Vasconcelos-Raposo, 2012), more frequently some form of multidimensionality was identified (e.g., Gao et al., 2012; Rey, Abad, Barrada, Garrido, & Ponsoda, 2014). The latter is sometimes interpreted as a methodological artifact resulting from wording effects because the GHQ-12 measures positive and negative self-appraisals with opposing keyed items (e.g., Hankins, 2008a). In contrast, others suggested that the GHQ-12 measures qualitatively different constructs such as general dysphoria and social dysfunction (Politi, Piccinelli, & Wilkinson, 1994) and, thus, allows for the interpretation of different subscale scores. Unfortunately, many of these findings are difficult to evaluate because they are based on highly selective samples or do not report the fit of competing models. Therefore, the present study examined the structure of the GHQ-12 within a meta-analytic structural equation modeling (MASEM) framework (Cheung & Hong, 2017; see also Gnambs & Staufenbiel, 2016). Two meta-analyses using either summary data or individual responses from multiple samples evaluated the dimensionality of the scale and scrutinized to what degree the GHQ-12 variance can be explained by a general factor or more dimensions.

Psychometric Properties of the General Health Questionnaire

The original GHQ consists of 60 items that were subsequently reduced to shorter versions with 30, 28, 20, or 12 items (Goldberg & Williams, 1988). The GHQ-12 includes six positively phrased items (e.g., "Have you been able to concentrate on what you were doing") and six negatively worded items (e.g. "Have you lost much sleep over worry") with four-point response scales (see Table 1). Standard Likert summation yields a global score between 0 and 36, a higher value reflecting more psychological distress (scoring method 0-1-2-3). However, other scoring schemes are also commonly used (see Rey et al., 2014), for example, a dichotomizing that collapses different response categories (0-0-1-1). Studies in different countries have reported a number of good psychometric properties of the GHQ-12 with respect to reliability and validity. Internal consistency reliabilities of the global score ranged from .79 to .91 (Hankins, 2008b; Shevlin & Adamson, 2005), whereas composite reliabilities approached .90 for different scoring methods (Rey et al., 2014). Moreover, test-retest reliabilities fell around .84 after 7 to 14 days (Piccinelli, Bisoffi, Bon, Cunico, & Tansella, 1993), at .79 after 20 days (López-Castedo & Fernández, 2005), and, as could be expected, declined with increasing retest-interval length, *r* = .68 after 12 weeks (Quek, Low, Razack, & Loh, 2001). Validities across 17 studies exhibited a median sensitivity of .84 and a median specificity of .79 (Goldberg et al., 1997). Similar values were found in subsequent studies (Martin & Newell, 2005). Construct validity was also established by means of convergent validity. As expected, the GHQ-12 global score showed a negative correlation with a global quality of life score (Montazeri et al., 2003) and positive associations with depression, state anxiety, and negative affectivity (Tait, French, & Hulse, 2003).

A more controversial issue concerns the factor structure underlying the GHQ-12. Originally, the GHQ-12 was designed as a unidimensional measure. Only a few studies corroborated this single factor structure (e.g., Fernandes & Vasconcelos-Raposo, 2012). More support exists for two- and three-dimensional models (e.g., Graetz, 1991; Martin, 1999; Politi et al., 1994). In an early study, Politi and colleagues (1994) identified two factors that were labelled 'General Dysphoria' and 'Social Dysfunction'. Whereas the latter included items relating to enjoying and coping with daily problems, the former reflected general anxiety and depression. Although the item factor correspondences were not always the same, similar results were found using exploratory (e.g., Iwata, Okuyama, Kawakami, & Saito, 1988; Schmitz, Kruse, & Tress, 1999) and confirmatory factor analyses (e.g., Gao et al., 2012; Gouveia, Barbosa, Andrade, & Carneiro, 2010). However, other two-factor models which are considerably different from the Politi et al. (1994) model have also been suggested (e.g., Li, Chung, Chui & Chan, 2009; Vanheule & Bogaerts, 2005). Using latent trait modeling and data from the GHQ-30, Andrich and van Schoubroeck (1989) demonstrated that positively and negatively worded items behave differently which can result in a methodological (artifactual) dual-factor model with all positively worded items loading on one factor and the negatively framed items on the other. Unfortunately, there is a large overlap between the substantively meaningful model of Politi and colleagues (1994) and the methodological artifact model. That is, the items constituting the 'General Dysphoria' factor are positively worded and the item of the 'Social Dysfunction' largely negatively (item 12 is assigned to both factors). This makes it difficult to disentangle the substantial psychological construct model from the methodological artifact model. In the literature, also alternative models with three factors have been proposed (e.g., Gao et al., 2004; Graetz, 1991; Martin, 1999; Shevlin & Adamson, 2005), some of which are quite diverse (Campbell, Walker, & Farrell, 2003). For example, Graetz (1991) found support for a 3-factor structure and distinguished between

‘Anxiety’ (comprising all positively worded items), ‘Anhedonia’, and ‘Loss of confidence’ (a breakdown of the negatively worded items in two factors).

Recently, studies tried to identify the most appropriate structure by comparing fit measures of a wider range of different models using confirmatory factor analyses. Here, too, no consistent structure emerged. For example, Tomás, Gutiérrez, and Sancho (2017) evaluated 20 different structural models of the GHQ-12. Among them were several rarely examined bifactor structures (cf. Reise, 2012) that allowed for facet-specific residual variations beyond a general factor common to all 12 items. These analyses supported Graetz’s (1991) 3-factorial model. In contrast, other studies found considerable support for alternative three factor models (e.g., Campbell, Walker, & Farrell, 2003) or even two factor models (e.g., Li, Chung, Chui & Chan, 2009; Rey et al., 2104; Vanheule & Bogaerts, 2005).

Taking into account that the multidimensionality of the GHQ-12 can (at least partially) be explained by artifactual wording effects, newer studies also tried to control for this effect statistically (Hankins, 2008a; Smith, Oluboyede, West, Hewison, & House, 2013; Ye, 2009; Wang & Lin, 2011). The control of this method bias was achieved in confirmatory factor models, which either allowed correlated errors or included additional method factors for differently worded items. In these studies, controlling for wording effects in the GHQ-12 showed a superior fit as compared with models not controlling for different item wording (e.g., Hankins, 2008a; Smith et al., 2013). Studies on the invariance of the factor structure across English and Chinese language versions of the GHQ-12 could also confirm the adequacy of the unidimensional model with wording effects (Chin et al., 2015). However, again the picture is not fully consistent. In some studies, the Graetz (1991) 3-factor model outperformed the single factor model, even when wording effects were controlled for (Abubakar & Fischer. 2012; Tomás et al., 2017). However, sometimes these results are difficult to compare because models were tested for variants of the GHQ-12 that removed some items (e.g., Wong & O'Driscoll, 2016) or introduced error covariances between items (e.g., Fernandes & Vasconcelos-Raposo, 2012). Additionally, the findings are further complicated by the use of different scoring methods that influence the factor structure and model fit (Rey et al., 2014).

Present Studies

The ongoing controversy surrounding the structure of the GHQ-12 led us to scrutinize its dimensionality from a meta-analytic perspective. Given the prevalent emphasis on replicability in psychological research (e.g., Open Science Collaboration, 2015), we sought to replicate our results in two independent meta-analyses using different data sources and different methodological approaches. Both meta-analyses adopted variants of MASEMs (Cheung & Hong, 2017) to derive pooled correlation matrices between the 12 items included in the GHQ-12 (see also Gnambs & Staufenbiel, 2016). Whereas the first meta-analysis relied on summary data and evaluated the structure of the GHQ-12 using an exploratory approach, the second meta-analysis adopted a confirmatory approach using individual responses from several samples. Moreover, bifactor modeling (Reise, 2012) allowed us to estimate the proportion of common variance explained by a general factor and, thus, to evaluate the meaningfulness of potential subscales. Despite several structural models that have been proposed in the literature (many of which differ on rather few parameters) no consensus as to the adequacy of these models has been reached. Therefore, we adopted meta-analytic exploratory factor analyses to evaluate the GHQ-12 without imposing zero-loading constraints on the loading matrix. This data-driven approach allowed us to evaluate potential (unhypothesized) cross-loadings on several factors. Meta-analytic confirmatory factor analyses were used to capture the multidimensionality of the scale implied by different structural models described in the literature. Although local misspecifications (e.g., missing factor loadings) can also be identified in confirmatory factor analyses, for example, using modification indices (Saris, Satorra, & van der Veld, 2009), these techniques haven not yet been evaluated within MASEM. Rather, MASEMs are typically judged by model-based goodness-of-fit indices which are known to be sensible to, among others, the item per factor ratio or the average factor loadings (Greiff & Heene, 2017). Therefore, the present study cross-validated the structure of the GHQ-12 in two complementary meta-analyses within an exploratory and a confirmatory framework.

Meta-Analysis I: Exploratory Analyses of Summary Data

Method

Meta-analytic database. Studies reporting on the factor structure of the GHQ-12 were identified in major academic (PsycINFO, Psyndex, EconLit, Business Source Complete, ERIC, SocINDEX, Medline, Scopus, Web of Science, ProQuest Dissertations & Theses Database) and non-academic databases (Google Scholar, Researchgate.net). Using the Boolean expression *general health questionnaire* AND (*exploratory factor analysis* OR *principal components* OR *correlation matrix*) these searches identified in October 2017 a total of 4,668 potential studies. After reviewing the title and the abstracts of these studies, 163 studies were further examined for inclusion in the meta-analytic database. Studies were retained according to the following criteria: (a) The study administered the 12 items included in the GHQ-12. We also considered longer versions of the GHQ as long as they subsumed all items of the GHQ-12. (b) The items were accompanied by their original four-point response scale and (c) used Likert coding of the responses (0-1-2-3). Because factor analyses of Pearson correlations among dichotomous variables typically result in distorted factor solutions (e.g., Kubinger, 2003), we did not include studies that adopted the dichotomous scoring method (0-0-1-1). (d) The study reported the results of an exploratory factor analysis or provided a full correlation matrix between the 12 items. (e) In case of oblique factor rotations, we only considered studies that also reported the respective factor correlations. (f) Moreover, we excluded factor pattern matrices with an excessive number of missing values (i.e., more than 50%). (g) Finally, one study (Gouveia et al., 2010) was excluded because it reported a nonpositive definite correlation matrix. The results of this search and screening process including a list of excluded studies are summarized in the supplemental material. In total, we identified 38 studies reporting on 45 independent samples that met our inclusion criteria.

Coding process. The authors developed a coding protocol (see supplemental material) for the extraction of relevant information from each publication that defined all variables and provided guidelines regarding the range of potential values. Two focal statistics were extracted from each study: If a study reported the correlations between the 12 items of the GHQ-12, we noted the respective correlation matrix. Otherwise, we retrieved the factor loadings and the respective factor correlations. In cases where different factor solutions were available for a given sample, we used the factor loading pattern including the largest number of factors. In addition, we extracted descriptive information on the sample (e.g., sample size, country, mean age, percentage of female participants), the publication (e.g., publication year), and the reported factor analysis (e.g., factor analytic method, type of rotation). All codings were conducted by the first author. To evaluate the coding process, 12 randomly selected studies (including about 30% of all samples) were independently coded a second time by a graduate student in psychology. For continuous variables (e.g., factor loadings) intercoder agreement was quantified using two-way intraclass coefficients (ICC; Shrout & Fleiss, 1979); for categorical variables (e.g., factor analytic method) we computed Cohen’s (1960) Kappa κ.

According to prevalent guidelines (see LeBreton & Senter, 2008) intercoder agreement can be considered strong for values exceeding .70 and excellent for values greater than .90. The intercoder reliability was ICC = .97, 95% CI [.966, .977] for the factor loadings and ICC = 1.00, 95% CI [1.00, 1.00] for the factor correlations. The remaining variables (e.g., sample size, factor analytic method) had ICCs or Cohen's κ of 1.00. The first author resolved disagreements by revisiting the respective study.

Meta-analytic procedure. The Pearson product-moment correlations between the 12 items of the GHQ-12 were used as effect size measures. Eleven samples reported the respective correlation matrix, whereas 34 samples reported only factor pattern matrices from exploratory factor analyses. For the latter, we calculated the implied correlations between the GHQ-12 items (see indirect method in Gnambs & Staufenbiel, 2016). In eight cases, only partial factor pattern matrices were available because small loadings (e.g., values falling below .40) were not reported. For these matrices, a value of 0 was imputed for the missing factor loadings. Monte Carlo simulations indicated that this approach results in unbiased estimates of the salient factor loadings (Gnambs & Staufenbiel, 2016).

The factor structure of the GHQ-12 was examined with a variant of two-step MASEM (see Cheung & Hong, 2017). In the first step, the item-level correlation matrices were pooled using a multivariate random-effects meta-analysis. Following Cheung (2013), we adopted a structural equation modeling (SEM) framework with a maximum likelihood estimator. In the second step, the thus derived pooled correlation matrix was submitted to an exploratory weighted least square factor analysis. As suggested by Cheung and Chan (2005), the asymptotic sampling covariance matrix of the pooled correlations was used as weight matrix for these analyses. In addition to a direct oblimin rotation (δ = 0; Bernaards & Jennrich, 2005), we also performed a target rotation to a partially specified bifactor structure (Browne,

1972) to disentangle scale-specific factors from a potential general factor underlying all items of the GHQ-12. A diverse set of criteria were used to decide on the number of factors to retain. These included eigenvalue-based criteria such as Kaiser's (1960) rule and Horn's (1965) parallel analysis, Velicer's (1976) minimum average partial test, as well as model fit indices such as the root mean square error of approximation (RMSEA; Browne & Cudeck, 1992). The robustness of the identified factor structure was evaluated in sensitivity analyses that repeated the meta-analysis within various subgroups. The similarity of the factor structures across these subgroups was quantified using coefficients of congruence for individual factors (Tucker, 1951) and coefficients of congruence for complete factor loading matrices (Gebhardt, 1968). Values between .85 and .94 indicate fair similarity, whereas factor structures with values of .95 or above can be considered equal (Lorenzo-Seva & ten Berge, 2006).

Statistical software and data availability. The correlations were pooled using the *metaSEM* software version 0.9.16 (Cheung, 2015) in *R* version 3.4.2. The factor analyses were conducted using routines based on the *psych* package version 1.7.8 (Revelle, 2017) and the *GPArotation* package version 2014-11-1 (Bernaards & Jennrich, 2005). To promote transparency and reproducibility of our analyses (see Nosek et al., 2015), all coded data and analyses scripts are provided in an online repository at http://osf.io/z5c4q/.

Results

Study characteristics. The meta-analysis included 45 independent samples that were published between 1983 and 2016 (*Mdn* = 2006). Each sample comprised of about *Mdn* = 446 participants (total *N* = 76,473; *Min* = 125; *Max* = 8,978) with approximately 54% women and a reported mean age of 36.87 years (*SD* = 16.04). The studies were conducted in 28 different countries around the world, with most samples coming from England (13%), Brazil (11%), and Japan (9%). The samples primarily administered the GHQ-12 (80%), whereas the rest received longer versions including either 20 items (4%) or 30 items (16%). The factor analyses of the GHQ-12 predominantly extracted two factors (84%); the remaining samples reported three factor solutions. The characteristics of each individual sample are also summarized in the supplemental material.

Meta-analytic factor analyses. The homogeneity of the correlation matrices was examined using a fixed-effects model. The respective fit indices (CFI = .80, RMSEA = .12, SRMR = .12) did not support the assumption of homogenous correlation matrices across samples. Therefore, we selected a random-effects model. The pooled correlations for the 12 items of the GHQ-12 (see supplemental material) ranged between .18 and .52 (*Mdn* = .30), whereas the respective random variances fell at *Mdn* = .009 (*Min* = .004, *Max* = .023). A diverse set of decision criteria suggested the extraction of two factors: (a) The first two unrotated eigenvalues exceeded 1 (λ_1 = 4.52 and λ_2 = 1.37), whereas the third did not (λ_3 = 0.81). (b) Velicer's (1976) minimum average partial criterion for one to four factor solutions fell at {.020, .020, .032, .049} and thus reached a minimum at one or two factors. (c) The RMSEA indicated a good model fit (i.e., a RMSEA < .05; Browne & Cudeck, 1992) for two factors, RMSEA_2 = .04, but not for a single factor, RMSEA_1 = .09. In contrast, Horn's (1965) parallel analysis suggested the extraction of three factors. Because most of these criteria pointed at two substantial factors, we conducted an exploratory factor analysis with oblique rotation extracting two factors. The respective results are summarized in Table 1. The two factors closely mirrored the multidimensional model introduced by Andrich and van Schoubroeck (1989) that separates the positively and negatively keyed items into distinct facets. On each factor six items had salient loadings, $M(|\lambda|)$ = .60, whereas the other items exhibited minor cross-loadings, $M(|\lambda|) = .08$. The two factors were substantially correlated at $r = .61$.

Bifactor modeling. Given the correlated factor structure, we examined to what degree the item variances could be explained by a general factor underlying all items of the GHQ-12. To this end, we conducted another exploratory factor analysis with an orthogonal target rotation toward a partially specified bifactor structure (Browne, 1972). The bifactor structure included a general factor for all items and two specific factors for the differently keyed items. The three latent factors were uncorrelated. The general factor can be interpreted as general distress, whereas the specific factors capture the residual variance due to the positively or negatively worded items. The respective results are summarized in Table 1. All items had loadings greater than .40 on the general factor, $M(|\lambda|) = .56$. In contrast, no item had salient loadings ($\lambda > .40$) on the specific factors, $M(|\lambda|) = .18$. Moreover, more than half of the common variance in each item was explained by the general factor (see last column in Table 1). Similarly, about 79% of the explained common variance was attributable to the general factor, whereas the specific factor for the positively and negatively worded items captured 16% and 5%, respectively. Thus, for a large part, the responses to the GHQ-12 were dominated by a single general factor.

Sensitivity analyses. The robustness of the identified factor structure was studied by repeating the meta-analytic bifactor analysis within various subgroups of samples and examining the similarity of the resulting factor structures. We selected three criteria and compared meta-analytic factor structures[1] derived from (a) reported correlation matrices (k = 11, N = 21,715), full factor loading matrices (k = 26, N = 43,068), and loading matrices with imputed missing values (k = 8, N = 11,690), (b) the GHQ-12 (k = 36, N = 61,932) and longer GHQ versions including either 20 or 30 items (k = 9, N = 14,541), and (c) English (k = 10, N

= 33,947), Spanish (k = 7, N = 4,972), Portuguese (k = 5, N = 8,713), and Japanese (k = 4, N = 6,036) language versions. The overall factor structures exhibited high similarity across these criteria (see supplemental material); the factor structure congruence coefficients fell between .96 and .99 (*Mdn* = .98). Particularly, the general factor was robustly replicated across the examined subgroups, *Mdn* = 1.00 (*Min* = .99, *Max* = 1.00); in contrast, the specific factors showed somewhat larger variability (*Min* = .79, *Max* = .99).

Meta-Analysis II: Confirmatory Analyses of Individual-Participant Data

The second meta-analysis extends the previous study on four central accounts: First, instead of summary statistics the present study focuses on individual responses of participants (see Debray et al., 2015). Thus, no potentially biasing reconstructions from incomplete factor loading matrices are necessary. Second, the study relied on participants from a single cultural and language group to avoid potential distortions resulting from imperfect test adaptations. Third, we used only representative samples from large-scale assessments to minimize sampling error and identify a common factor pattern for a given population. Fourth, the previously identified factor structure of the GHQ-12 was tested using a confirmatory approach. Thus, the study intends to replicate the previous results in an individual-participant meta-analysis using a new data source and adopting a different analytical approach.

Method

Meta-analytic database. Individual participant data for the GHQ-12 were retrieved from the *UK Data Archive* (http://www.ukdataservice.ac.uk), a non-profit data catalogue for social, health, and economic surveys conducted in the United Kingdom, using the search term *general health questionnaire*. A sample was included in the meta-analysis if it (a) administered the 12 items of the GHQ-12, (b) in its English language version, (c) accompanied by their original four-point response scales, and (d) drew a representative sample from the population of the United Kingdom or one of its countries. This search process identified 84 independent samples from several large-scale health and social surveys in England, Scotland, and Northern Ireland. A full list of all included samples is given in the supplemental material.

Meta-analytic procedure. The structure of the GHQ-12 was evaluated by two-step MASEM (Cheung & Hong, 2017). In the first step, we calculated the correlation matrix for the 12 items within each sample (see Cheung & Jak, 2016). Negatively worded items were reverse coded. The correlation matrices for each sample are available at http://osf.io/z5c4q/. As in the previous meta-analysis, these correlation matrices were pooled across samples using SEM with maximum likelihood estimation. In the second step, several confirmatory factor models were fitted to the pooled correlation matrix using a weighted least square estimator. Again, the asymptotic sampling covariance matrix of the pooled correlations was used as weight matrix for these analyses (Cheung & Chan, 2005). Both analyses steps were conducted with the *metaSEM* software version 0.9.16 (Cheung, 2015). The fit of these models was evaluated in line with conventional criteria (cf. Schermelleh-Engel, Moosbrugger, & Müller, 2003) using the *Comparative Fit Index* (CFI), the *Standardized Root Mean Square Residual* (SRMR), and the RMSEA. Models with a CFI ≥ .95, a RMSEA ≤ .08, and a SRMR ≤ .10 were interpreted as "acceptable" and models with CFI ≥ .97, RMSEA ≤ .05, and SRMR ≤ .05 as "good" fitting.

Examined factor models. Different structural models were evaluated that have been frequently used in previous research. All models included unconstrained factor loadings and uncorrelated item uniquenesses. The latent factor variances were fixed to 1 for model identification.

In line with the original construction rationale of the GHQ-12 (Goldberg, 1972), *Model 1* included a single factor explaining the covariances between all items. In contrast, *Model 2* additionally acknowledged potential wordings effects (see Ye, 2009; Wang & Lin, 2011). Thus, we estimated a general factor for all items and an orthogonal specific factor for the negatively worded items (see Figure 1). Sometimes, these types of models are also termed nested factor models (Schulze, 2005) or bifactor-(*S*-1) models (Eid, Geiser, Koch, & Heene, 2017). *Model 3* followed Andrich and van Schoubroeck (1989) and specified two correlated latent factors for the positively (1, 3, 4, 7, 8, 12) and negatively worded items (2, 5, 6, 9, 10, 11; see Figure 1). These factors have either been interpreted as representing wording effects (Hankins, 2008ab) or qualitatively different types of mental health, general dysphoria and social dysfunction (Politi et al., 1994). Because these factors were typically correlated, we also estimated a bifactor structure (see Reise, 2012) to disentangle the effects of a general factor from specific factor influences. Thus, we modeled a general factor common to all 12 items and two orthogonal specific factors for the differently worded items (Model 3b in Figure 1). *Model 4* was introduced by Graetz (1991) and included three correlated factors representing anxiety (2, 5, 6, 9), social dysfunction (1, 3, 4, 7, 8, 12), and loss of confidence (10, 11). An alternative three-factor solution was suggested by Martin (1999). Thus, *Model 5* specified three correlated factors reflecting depression (6, 9, 10, 11, 12), stress (2, 5, 7), and successful coping (1, 3, 4, 8). Again, these models were also estimated with a bifactor structure to separate general and specific factor effects.

Results

The meta-analysis included 84 independent samples that were surveyed between 1987 and 2013 (*Mdn* = 2003). These samples included *N* = 410,640 participants (57% women) in the age from 13 to 100 years (*M* = 45.05; *SD* = 19.55). The ICCs for all items of the GHQ-12 were very small (all ICCs ≤ .014). Thus, most of the variance in the observed item scores was a result of individual differences between participants and not between samples. Accordingly, meta-analytic models including random effects for the correlations between the 12 items did not converge. In contrast, a fixed-effects model indicated a satisfactory fit (CFI = .99, RMSEA = .03, SRMR = .04) and, thus, homogenous correlation matrices across samples. The pooled correlation matrix of the GHQ-12 is given in the supplemental material. The correlations ranged from .22 to .68 (*Mdn* = .38).

Meta-analytic factor analyses. The structure of the GHQ-12 was examined by fitting different confirmatory models to the pooled correlation matrix. The respective fit statistics in Table 2 corroborated the findings of our first meta-analysis. The unidimensional model clearly exhibited an unsatisfactory fit (CFI = .89, SRMR = .15, RMSEA = .11). In contrast, most multidimensional models showed at least acceptable fits. However, the oblique three-factor model suggested by Martin (1999) demonstrated an inferior fit as compared to the other models; the respective bifactor formulation even failed to converge. The best fit in terms of the information criteria represented the bifactor formulation of Andrich and van Schoubroeck (1989) that acknowledged different wording effects (CFI = .97, SRMR = .04, RMSEA = .05). The standardized factor loadings are given in Figure 1. All items had loadings greater than .40 on the general factor, $M(|\lambda|) = .63$ (*Min* = .42, *Max* = .84). But, also the specific factor for the positively worded items exhibited substantial loadings, $M(|\lambda|) = .44$ (*Min* = .32, *Max* = .55). In contrast, the specific factor for the negatively worded items had a rather unclear loading pattern, $M(|\lambda|) = .22$ (*Min* = .07, *Max* = .39). Together, the two specific factors explained about 23% of the common variance, whereas most of the explained common variance (77%) was attributable to the general factor (see Table 3). The total score reliability, that is, the proportion of variance in GHQ-12 scores accounted for by the general factor, was $\omega_H = .85$.

To evaluate the meaningfulness of subscale scores in the GHQ-12, we also calculated omega hierarchical subscale ($\omega_{H.S}$; see Rodriguez, Reise, & Haviland, 2016) for the positively and negatively keyed items. This reflects the proportion of unique variance in the subscale scores (reliability) after accounting for the general factor. $\omega_{H.S}$ was estimated as .39 for the positively worded items and .01 for the negatively worded items indicating that both subscales reflected negligible unique variance and primarily represented the general factor. Thus, in line with the previous meta-analyses, the responses to the GHQ-12 seemed to be dominated by a single general factor.

Replicability of meta-analytic factor structure. The robustness of the identified factor solution was evaluated by comparing the pooled correlation matrices from the two meta-analyses. On average, the pooled correlations derived in the second meta-analysis were all larger than the respective correlations from the first meta-analysis, $M(\Delta r)$ = .09 (SD = .04). However, as summarized in Table 2, the confirmatory factor analyses fitted to the pooled correlation matrix from the first meta-analysis replicated the previously reported results: Multidimensional models outperformed the single factor model; albeit, again the Martin (1999) model provided an inferior fit. Moreover, bifactor specifications fitted better than comparable correlated trait models. Again, the GHQ-12 was dominated by a general factor explaining between 69% and 75% of the common variance (see Table 3), whereas specific factors were less clearly represented (2% to 30%). Subgroup analyses for different language versions replicated these results (see supplemental material). Finally, a multi-group analysis for the bifactor model of Andrich and van Schoubroeck (1989) showed configural measurement invariance across the two meta-analyses (χ^2 = 43,650, df = 84, CFI = .0971, RMSEA = .033). Placing equality constraints on the factor loadings did not result in a drop of the CFI below what is usually considered acceptable ($\Delta\chi^2$ = 1,735, Δdf = 12, ΔCFI < .002;

Meade, Johnson, & Braddy, 2008; Khojasteh & Lo, 2015) and, thus, confirmed metric measurement invariance.

Discussion

For decades, the GHQ-12 has dominated mental health screenings in applied research and clinical practice (Fryers et al., 2004). It is surprising that after more than 40 years a fundamental debate on the structure of the GHQ-12 has not been resolved. Empirical studies frequently identified unidimensional as well as various multidimensional factor solutions (e.g., Fernandes & Vasconcelos-Raposo, 2012; Gao et al., 2012; Rey et al., 2014). However, different sample characteristics, language versions, and analyses methods adopted in these studies made it difficult to find a consensus. To reconcile these conflicting results, we presented two meta-analyses that systematically examined the factor structure of the GHQ-12 across samples. These analyses provided three central findings. First, the GHQ-12 is not strictly unidimensional but also reflects wording effects (see also Hankins, 2008a). Particularly, positively keyed items explained incremental variance beyond a general mental health factor. Thus, latent variable modeling needs to acknowledge these dependencies to properly account for the covariance structure of the GHQ-12. Second, the bifactor structure with wording effects was rather robust and replicated across different language versions. Particularly, the general factor and the specific factor pertaining to positively worded items were highly similar across English, Spanish, Portuguese, and Japanese translations of the instrument. In contrast, the specific factor for negatively worded items showed more variability across language versions. Thus, negatively worded items seem to reflect some form of language-specific variance such as cross-cultural differences in response styles (Johnson, Kulesa, Cho, & Shavitt, 2005). Finally, bifactor modeling revealed that a single dominant factor accounted for most of the item variance. In contrast, subscale-specific variance associated with the wording of the items was rather negligible. Overall, the GHQ-12 seems to represent an essentially unidimensional instrument with spurious secondary dimensions reflecting the wording of the items.

Implications for Applied Measurement

It is not uncommon for many psychological measures used in applied practice to capture a dominant general factor, while also reflecting some minor secondary dimensions (Reise, Moore, & Haviland, 2010) that capture, for example, facet-specific variance (e.g., Henry & Crawford, 2005; Vasconcelos-Raposo, Fernandes, & Teixeira, 2013) or systematic response styles (e.g., Marsh, 1996). Similar, the GHQ-12 is not strictly unidimensional, but also reflects systematic residual variance beyond a general distress (or, reverse coded, mental health) factor. Because these residual variances pertained to differently worded items, this pattern can be interpreted as an expression of specific response styles such as acquiescence (Hankins, 2008a). Thus, the multidimensionality of the GHQ-12 seems to reflect method-specific variance that needs to be controlled for in latent variable analyses modeling responses to the 12 items. In practice, pronounced multidimensionality is problematic if composite scores (e.g., sum scores across all items) are used because these reflect a blend of different latent traits. However, for the GHQ-12 these secondary dimensions seem to be less influential; more than 75% of the explained variance was attributable to the general factor. Thus, applied researchers are likely to introduce a negligible bias in their analyses if they adopt composite scores and ignore wording effects. On the other hand, these results also cast doubts on the usefulness of subscale scores calculated separately for negatively and positively worded items (or other facet models; e.g., Graetz, 1991). In our analyses, respective subscales were highly correlated (r = .80), and, thus shared a large proportion of variance. This was also reflected in rather low reliability estimates showing rather limited unique variance captured by subscales for negatively and positively worded items. Because subscales in the GHQ-12 primarily reflect general factor variance and to a lesser degree unique variance, it does not seem advisable to use these scores in substantial analyses. Indeed, comparative analyses showed that subscale scores rarely exhibited substantially different associations with criterion variables as compared to composite scores for the entire scale (e.g., Aguado et al., 2012; Gao et al., 2004; Shevlin & Adamson, 2005). Researchers interested in a more fine-grained differentiation of mental health would likely be better served with longer versions of the GHQ that exhibit clearer facet structures (see, for example, Klainin-Yobas & He, 2014, on the GHQ-30) or alternative instruments such as the Short-Form 36 (SF-36) Health Survey (Anagnostopoulos, Niakas, & Tountas, 2009).

Cautionary Notes and Outlook

The present studies relied on summary statistics pooled across multiple independent samples to scrutinize the structure of the GHQ-12. Accordingly, these results refer to the factor structure in an average sample (in terms of, for example, sociodemographic, cultural, or psychological characteristics of the included respondents). It is conceivable that specific sample characteristics such as individual differences in reading competences (Gnambs & Schroeders, 2017) or random responding (Huang, Liu, & Bowling, 2015) might contribute to ambiguous factor structures in a given sample that deviates from the presented results. Therefore, future research is encouraged to identify moderating influences that might contribute to the multidimensionality in psychological measures. For example, our results were limited to the Likert scoring method of the GHQ-12 and do not necessarily extend to different scoring schemes (see Rey et al., 2014). Similar, group comparisons require a coherent measurement of mental health across, for example, different assessment contexts (e.g., paper versus computerized tests; cf. Gnambs & Kaspar, 2017), measurement occasions

(Mäkikangas et al., 2006), cultural settings (Romppel et al, 2017), or respondent groups (e.g., clinical versus community samples). Moreover, we also want to emphasize that the comparable factor structure identified in different language versions of the GHQ-12, does not relieve researchers from demonstrating measurement invariance in the specific sample at hand. Finally, we hope to see more research that demonstrates the incremental validity of potential subscale scores beyond a general factor before using and interpreting these scales.

Conclusion

Although the GHQ-12 is not strictly unidimensional, specific factors associated with the item wording explain limited and contrasting unique variance beyond a general factor. Therefore, composite scores are likely to exhibit only a minor bias resulting from ignored multidimensionality. In contrast, it is not recommended to use and interpret subscale scores because they primarily reflect general mental health rather than distinct constructs.

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

Anagnostopoulos, F., Niakas, D., & Tountas, Y. (2009). Comparison between exploratory factor-analytic and SEM-based approaches to constructing SF-36 summary scores. *Quality of Life Research, 18*, 53-63. doi:10.1007/s11136-008-9423-5

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

Bernaards, C. A., & Jennrich, R. I. (2005). Gradient projection algorithms and software for arbitrary rotation criteria in factor analysis. *Educational and Psychological Measurement, 65*, 676-696. doi:10.1177/0013164404272507

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods, 1*, 97-111. doi:10.1002/jrsm.12

Browne, M. W. (1972). Orthogonal rotation to a partially specified target. *British Journal of Mathematical and Statistical Psychology*, *25*, 115-120. doi:10.1111/j.2044-8317.1972.tb00482.x

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods and Research, 21*, 230-258. doi:10.1177/0049124192021002005

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

Chin, E. G., Drescher, C. F., Trent, L. R., Darden, M., Seak, W. C., & Johnson, L. R. (2015). Searching for a screener: Examination of the factor structure of the General Health Questionnaire in Malaysia. *International Perspectives in Psychology: Research, Practice, Consultation*, *4*, 111-127. doi:10.1037/ipp0000030

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46. doi:10.1177/001316446002000104

Cheung, M. W.-L. (2013). Multivariate meta-analysis as structural equation models. *Structural Equation Modeling, 20*, 429-454. doi:10.1080/10705511.2013.797827

Cheung, M. W. L. (2015). metaSEM: An R package for meta-analysis using structural equation modeling. *Frontiers in Psychology, 5*. doi:10.3389/fpsyg.2014.01521

Cheung, M. W.-L., & Chan, W. (2005). Meta-analytic structural equation modeling: A two-stage approach. *Psychological Methods, 10*, 40-64. doi:10.1037/1082-989X.10.1.40

Cheung, M. W. L., & Hong, R. Y. (2017). Applications of meta-analytic structural equation modeling in health psychology: Examples, issues, and recommendations. *Health Psychology Review, 11*, 265-279. doi:10.1080/17437199.2017.1343678

Cheung, M. W. L., & Jak, S. (2016). Analyzing big data in psychology: a split/analyze/meta-analyze approach. *Frontiers in Psychology, 7*. doi:10.3389/fpsyg.2016.00738

Debray, T., Moons, K. G., Valkenhoef, G., Efthimiou, O., Hummel, N., Groenwold, R. H., & Reitsma, J. B. (2015). Get real in individual participant data (PID) meta-analysis: a review of the methodology. *Research Synthesis Methods*, *6*, 293-309. doi:10.1002/jrsm.1160

Eid, M., Geiser, C., Koch, T., & Heene, M. (2017). Anomalous results in g-factor models: Explanations and alternatives. *Psychological Methods, 22*, 541-562. doi:10.1037/met0000083

Fernandes, H. M., & Vasconcelos-Raposo, J. (2012). Factorial validity and invariance of the GHQ-12 among clinical and nonclinical samples. *Assessment*, *20*, 219-229. doi:10.1177/1073191112465768

Fryers, T., Brugha, T., Morgan, Z., Smith, J., Hill, T., Carta, M., ... & Kovess, V. (2004). Prevalence of psychiatric disorder in Europe: the potential and reality of meta-analysis. *Social Psychiatry and Psychiatric Epidemiology*, *39*, 899-905. doi:10.1007/s00127-004-0875-9

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

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: Dimensionality and item quality. *Psycho-Oncology*, *21*, 954-961. doi:10.1002/pon.1989

Gebhardt, F. (1968). Über die Ähnlichkeit von Faktormatrizen [On the similarity of factor matrices]. *Psychologische Beiträge, 10*, 591-599.

Gnambs, T., & Kaspar, K. (2017). Socially desirable responding in web-based questionnaires: A meta-analytic review of the candor hypothesis. *Assessment, 24*, 746-762. doi:10.1177/1073191115624547

Gnambs, T., & Schroeders, U. (2017). Cognitive abilities explain wording effects in the Rosenberg Self-Esteem Scale. *Assessment*. Advance online publication. doi:10.1177/1073191117746503

Gnambs, T., & Staufenbiel, T. (2016). Parameter accuracy in meta-analyses of factor structures. *Research Synthesis Methods, 7*, 168-186. doi:10.1002/jrsm.1190

Goldberg, D. P. (1972). *The detection of psychiatric illness by questionnaire*. London, England: Oxford University Press.

Goldberg, D. P, Gater, R., Sartorius, N., Ustun, T. B., Piccinelli, M., Gureje, O., & Rutter, C. (1997). The validity of two versions of the GHQ in the WHO study of mental illness in general health care. *Psychological Medicine*, *27*, 191-197. doi:10.1017/S0033291796004242

Goldberg, D. P., & Williams, P. (1988*). A users's guide to the General Health Questionnaire*. London, England: GL Assessment.

Gouveia, V. V., Barbosa, G. A., Andrade, E. O., & Carneiro, M. B. (2010). Factorial validity and reliability of the General Health Questionnaire (GHQ-12) in the Brazilian physician population. *Cadernos de Saúde Pública, 26*, 1439-1445. doi:10.1590/S0102-311X2010000700023

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

Greiff, S., & Heene, M. (2017). Why psychological assessment needs to start worrying about model fit. *European Journal of Psychological Assessment, 33*, 313-317. doi:10.1027/1015-5759/a000450

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). 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). doi:10.1186/1471-2458-8-355.

Henry, J. D., & Crawford, J. R. (2005). The short-form version of the Depression Anxiety Stress Scales (DASS-21): Construct validity and normative data in a large non-clinical sample. *British Journal of Clinical Psychology, 44*, 227-239. doi:10.1348/014466505X29657

Horn, J. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*, 179-185. doi:10.1007/BF02289447

Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology, 100*, 828-845. doi:10.1037/a0038510

Iwata, N., Okuyama, Y., Kawakami, Y., & Saito, K. (1988). The twelve-item General Health Questionnaire among Japanese workers. *Environmental Science, 11*, 1-10.

Johnson, T., Kulesa, P., Cho, Y. I., & Shavitt, S. (2005). The relation between culture and response styles: Evidence from 19 countries. *Journal of Cross-Cultural Psychology, 36*, 264-277. doi:10.1177/0022022104272905

Kaiser, H. F. (1960). The application of the electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 141-151. doi:10.1177/001316446002000116

Khojasteh, K., & Lo, W.-J. (2015). Investigating the sensitivity of goodness-of-fit indices to detect measurement invariance in a bifactor model. *Structural Equation Modeling, 22*, 531-541. doi:10.1080/10705511.2014.937791

Klainin-Yobas, P., & He, H. G. (2014). Testing psychometric properties of the 30-item general health questionnaire. *Western Journal of Nursing Research, 36*, 117-134. doi:10.1177/0193945913485649

Kubinger, K. D. (2003). On artificial results due to using factor analysis for dichotomous variables. *Psychology Science, 45*, 106-110.

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11*, 815-852. doi:10.1177/1094428106296642

Li, W. H. C., Chung, J. O. K., Chui, M. M. L., & Chan, P. S. L. (2009). Factorial 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

López-Castedo, A., & Fernández, L. (2005). Psychometric properties of the Spanish version of the 12-item General Health Questionnaire in adolescents. *Perceptual and Motor Skills, 100*, 676-680. doi:10.2466/pms.100.3.676-680

Lorenzo-Seva, U., & ten Berge, J. M. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology, 2*, 57-64. doi:10.1027/1614-2241.2.2.57

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

Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology*, *70*, 810-819. doi:10.1037/0022-3514.70.4.810

Martin, A. J. (1999). Assessing the multidimensionality of the 12-item General Health Questionnaire. *Psychological Reports, 84*, 927-935. doi:10.2466/pr0.1999.84.3.927

Martin, C. R., & Newell, R. J. (2005). Is the 12-item General Health Questionnaire (GHQ-12) confounded by scoring method in individuals with facial disfigurement? *Psychology and Health*, *20*, 651-659. doi:10.1080/14768320500060061

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology, 93*, 568-592. doi:10.1037/0021-9010.93.3.568

Montazeri, A., Harirchi, A. M., Shariati, M., Garmaroudi, G., Ebadi, M., & Fateh, A. (2003). The 12-item General Health Questionnaire (GHQ-12): Translation and validation study of the Iranian version. *Health and Quality of Life Outcomes*, 1:66. doi:10.1186/1477-7525-1-66

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., … Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*, 1420-1422. doi:10.1126/science.aab2374

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. doi:10.1126/science.aac4716

Piccinelli, M., Bisoffi, G., Bon, M. G., Cunico, L., & Tansella, M. (1993). Validity and test-retest reliability of the Italian version of the 12-item General Health Questionnaire in general practice: A comparison between three scoring methods. *Comprehensive Psychiatry*, *34*, 198-205. doi:10.1016/0010-440X(93)90048-9

Politi, P. L., Piccinelli, M., & Wilkinson, G. (1994). Reliability, validity and factor structure of the 12-item General Health Questionnaire among young males in Italy. *Acta Psychiatrica Scandinavica*, *90*, 432-437. doi:10.1111/j.1600-0447.1994.tb01620.x

Quek, K. F., Low, W. Y., Razack, A. H., & Loh, C. S. (2001). Reliability and validity of the General Health Questionnaire (GHQ-12) among urological patients: A Malaysian study. *Psychiatry and Clinical Neurosciences, 55*, 509-513. doi:10.1046/j.1440-1819.2001.00897.x

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47*, 667-696. doi:10.1080/00273171.2012.715555

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment*, *92*, 544-559. doi:10.1080/00223891.2010.496477

Revelle, W. (2017). *psych: Procedures for personality and psychological research*. Northwestern University, Evanston, IL. Retrieved from http://CRAN.R-project.org/package=psych

Rey, J. J., Abad, F. J., Barrada, J. R., Garrido, L. E., & Ponsoda, V. (2014). The impact of ambiguous response categories on the factor structure of the GHQ-12. *Psychological Assessment*, *26*, 1021-1030. doi:10.1037/a0036468

Romppel, M., Hinz, A., Finck, C., Young, J., Brähler, E., & Glaesmer, H. (2017). Cross-cultural measurement invariance of the General Health Questionnaire-12 in a German and a Colombian population sample. *International Journal of Methods in Psychiatric Research*. Advance online publication. doi:10.1002/mpr.1532

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, *21*, 137-150. doi:10.1037/met0000045

Saris, W. E., Satorra, A. & van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications. *Structural Equation Modeling, 16*, 561-582. doi:10.1080/10705510903203433

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

Schmitz, N., Kruse, J., & Tress, W. (1999). Psychometric properties of the General Health Questionnaire (GHQ-12) in a German primary care sample. *Acta Psychiatrica Scandinavica*, *100*, 462-468. doi:10.1111/j.1600-0447.1999.tb10898.x

Schulze, R. (2005). Modeling structures of intelligence. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 241-263). Thousand Oaks, CA: Sage Publications.

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

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420-428. doi:10.1037/0033-2909.86.2.420

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

Tait, R. J., French, D. J., & Hulse, G. K. (2003). Validity and psychometric properties of the General Health Questionnaire-12 in young Australian adolescents. *Australian and New Zealand Journal of Psychiatry*, *37*, 374-381. doi:10.1046/j.1440-1614.2003.01133.x

Tomás, J. M., Gutiérrez, M., & Sancho, P. (2017). Factorial validity of the General Health Questionnaire 12 in an Angolan sample. *European Journal of Psychological Assessment, 33*, 116-112. doi:10.1027/1015-5759/a000278

Tucker, L. R. (1951). *A method for synthesis of factor analysis studies* (Personnel Research Section Report No. 984). Washington, DC: Department of the Army.

Vanheule, S., & Bogaerts, S. (2005). The factorial structure of the GHQ-12. *Stress and Health*, *21*, 217-222. doi:10.1002/smi.1058

Vasconcelos-Raposo, J., Fernandes, H. M., & Teixeira, C. M. (2013). Factor structure and reliability of the depression, anxiety and stress scales in a large Portuguese community sample. *Spanish Journal of Psychology, 16*, 1-10. doi:10.1017/sjp.2013.15

Velicer, W. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika, 41*, 321-327. doi:10.1007/BF02293557

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

Wong, K. C. K., & O'Driscoll, M. P. (2016). Psychometric properties of the General Health Questionnaire-12 in a sample of Hong Kong employees. *Psychology, Health & Medicine*. doi:10.1080/13548506.2016.1140901

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

Footnotes

1) Random-effects meta-analyses using a small number of samples can result in unstable estimates of between-studies variances (Borenstein, Hedges, Higgins, & Rothstein, 2010). Accordingly, for some subgroup analyses respective random-effects model did not converge and did not give meaningful heterogeneity estimates for several pooled correlations. Therefore, subgroup analyses pertaining to correlation matrices as effects sizes as well as different language versions were based on a fixed-effects model, whereas all other analyses adopted a random-effects model.

Table 1.

Exploratory Factor Loading Patterns in Meta-Analysis I

		Single factor model		*Oblique factor model*			*Bifactor model*				
	Item	Factor 1	h^2	Factor 1	Factor 2	h^2	General factor	Specific factor 1	Specific factor 2	h^2	I-ECV
I01	Been able to concentrate on whatever you are doing	**.51**	.26	.17	**.41**	.28	**.49**	.12	.25	.32	.78
I03	Felt that you are playing a useful part in things	**.49**	.24	-.04	**.63**	.37	**.47**	-.11	.37	.37	.59
I04	Felt capable of making decisions about things	**.49**	.24	-.06	**.65**	.38	**.46**	-.07	.38	.37	.58
I07	Enjoyed normal day-to-day activities	**.54**	.29	.13	**.49**	.34	**.52**	.07	.29	.36	.76
I08	Been able to face up to your problems	**.52**	.27	.02	**.60**	.37	**.50**	-.10	.33	.36	.67
I12	Felt reasonably happy, all things considered	**.51**	.26	.09	**.50**	.32	**.49**	-.02	.28	.32	.75
I02	Lost much sleep over worry	**.54**	.29	**.65**	-.07	.37	**.56**	.27	-.17	.41	.78
I05	Felt constantly under strain	**.61**	.37	**.73**	-.07	.48	**.63**	.25	-.19	.50	.79
I06	Felt you could not overcome your difficulties	**.61**	.37	**.66**	.01	.44	**.65**	.06	-.17	.44	.94
I09	Been feeling unhappy and depressed	**.68**	.47	**.72**	.04	.55	**.72**	.08	-.16	.54	.94
I10	Been losing confidence in yourself	**.66**	.43	**.62**	.11	.47	**.70**	-.16	-.16	.53	.87
I11	Thinking of yourself as a worthless person	**.57**	.33	**.50**	.13	.34	**.60**	-.23	-.13	.44	.85
	Eigenvalue	3.81		2.71	2.01		3.92	0.27	0.78		
	Proportion of variance	32%		23%	17%		33%	2%	6%		
	Proportion of explained variance	100%		57%	43%		79%	5%	16%		

Note. Exploratory weighted least square factor analysis with direct oblimin (Bernaards & Jennrich, 2005) or target rotation (Browne, 1972). The factor correlation in the oblique case was $r = .61$. All items were recoded in such a way that higher values indicate better mental health. Factor loadings $|\lambda| \geq .40$ are in bold. h^2 = Communality; I-ECV = Proportion of common variance explained by the general factor (Rodriguez et al., 2016).

Table 2.

Fit Statistics for Different Confirmatory Factor Models of the GHQ-12.

		Meta-Analysis I							*Meta-Analysis II*						
	Model	χ 2	*f*	FI	S RMR	R MSEA	IC	IC	χ^2	*f*	FI	S RMR	R MSEA	A IC	B IC
.	Single factor model	1 ,007*	4	939	. 066	.0 15	99	00	1 59,642*	4	894	. 145	.0 84	1 59,534	1 58,944
.	Artifactual model	1 73*	8	992	. 027	.0 06	7	366	9 6,328*	8	936	. 086	.0 70	9 6,232	9 5,707
	Andrich & van Schoubroeck (1989)														
a.	Correlated factor model	2 07*	3	990	. 029	.0 06	01	389	9 8,620*	3	935	. 088	.0 67	9 8,514	9 7,935
b.	Bifactor model	7 5*	2	998	. 016	.0 03	9	397	4 3,575*	2	971	. 040	.0 50	4 3,491	4 3,032
	Graetz (1991)														
a.	Correlated factor model	1 57*	1	993	. 025	.0 05	5	416	7 8,832*	1	948	. 072	.0 61	7 8,730	7 8,173
b.	Bifactor model	7 5*	3	998	. 016	.0 03	11	408	4 3,705*	3	971	. 040	.0 50	4 3,619	4 3,149
	Martin (1999)														
a.	Correlated factor model	7 16*	1	958	. 056	.0 13	14	42	1 36,571*	1	910	. 123	.0 81	1 36,469	1 35,912
b.	Bifactor model	4 94*	2	971	. 046	.0 12	10	2	Model did not converge						

Note. *N* = 76,473 and 410,640 for meta-analyses I and II. CFI = Comparative Fit Index; SRMR = Standardized Root Mean Residual; RMSEA = Root Mean Square Error of Approximation; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion.

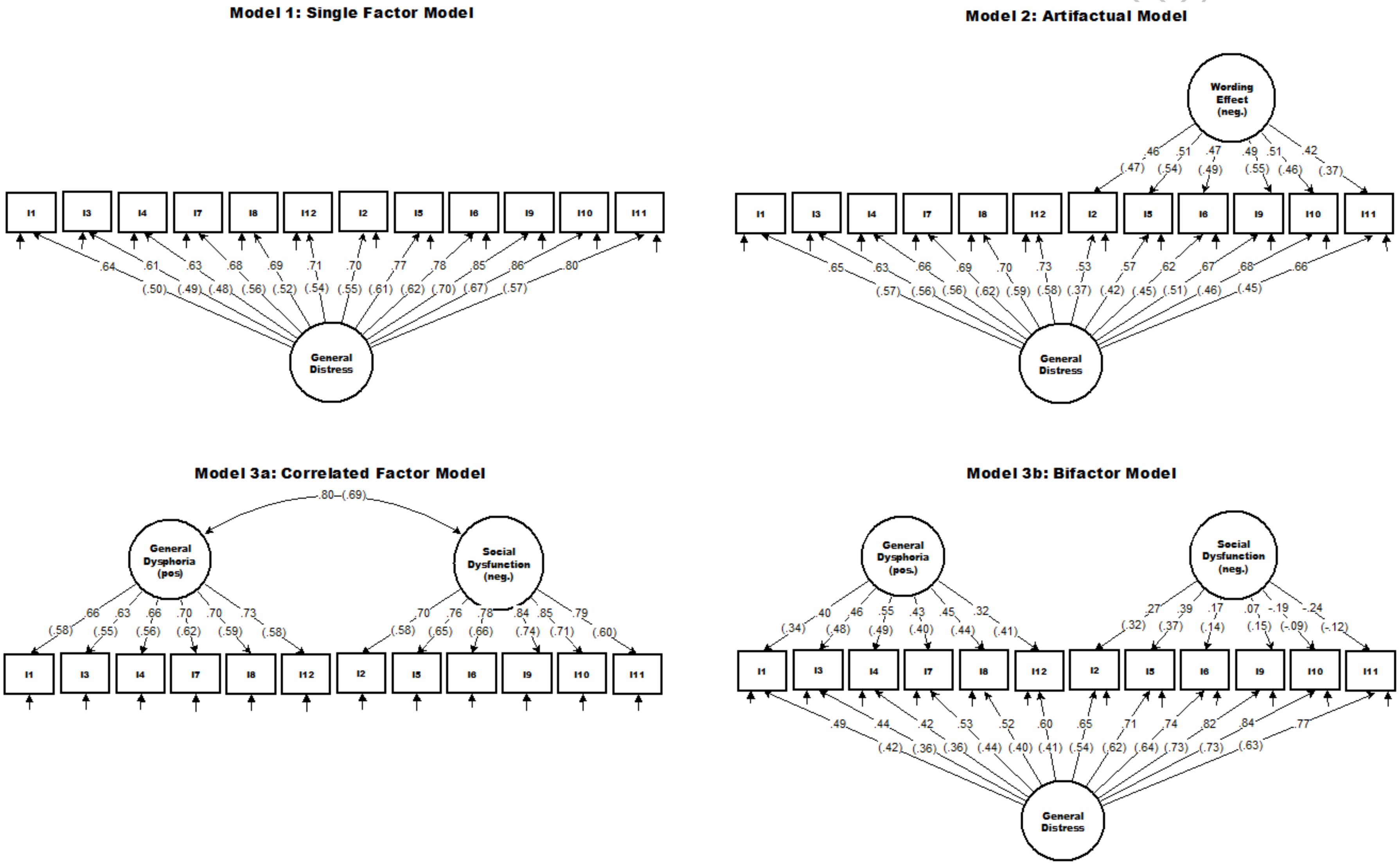
* $p < .05$

Table 3.

Score Reliabilities for Different General Factor Models of the GHQ-12.

Model	ω_h	ω_{s_1}	ω_{s_2}	ω_{s_3}	ECV_g	ECV_{s_1}	ECV_{s_2}	ECV
Meta-Analysis I								
Single factor model	.89							
Artifactual model	.75	.47			.70	.30		
Andrich & van Schoubroeck (1989)	.78	.43	.03		.72	.22	.06	
Graetz (1991)	.77	.41	.15	.08	.69	.21	.07	.02
Martin (1999)	.86	.31	.06	.06	.75	.12	.05	.08
Meta-Analysis II								
Single factor model	.96							
Artifactual model	.85	.36			.79	.21		
Andrich & van Schoubroeck (1989)	.85	.39	.01		.77	.18	.05	
Graetz (1991)	.84	.35	.15	.13	.74	.16	.07	.03

Note. ω_h = General factor reliability (i.e., proportion of variance in total scores attributed to the general factor); ω_s = Specific factor reliability (i.e., proportion of variance in subscale scores attributed to the specific factor); ECV = Proportion of common variance explained by the general / specific factor (see Rodriguez et al., 2016); s_1 = Items 1, 3, 4, 7, 8, and 12 (Models 2 and 3), 2, 5, 6, and 9 (Model 4), or 6, 9, 10, 11, and 12 (Model 5); s_2 = Items 2, 5, 6, 9, 10, and 11 (Model 3), 1, 3, 4, 7, 8, and 12 (Model 4), or 2, 5, and 7 (Model 5); s_3 = Items 10 and 11 (Model 4) or 1, 3, 4, and 8 (Model 5).



Model 1: Single Factor Model

Model 2: Artifactual Model

Model 3a: Correlated Factor Model

Model 3b: Bifactor Model

Figure 1. Factor models for the GHQ-12 with standardized parameter estimates for Meta-analysis I (in parentheses) and II.