Can the 12-item General Health Questionnaire be used to measure positive mental health?

YONGJIAN HU1, SARAH STEWART-BROWN1, LIZ TWIGG2 AND SCOTT WEICH1*

1 Health Sciences Research Institute, Warwick Medical School, University of Warwick, Coventry, UK; 2 Institute for the Geography of Health, Department of Geography, University of Portsmouth, Portsmouth, UK

ABSTRACT

Background. Well-being is an important determinant of health and social outcomes. Measures of positive mental health states are needed for population-based research. The 12-item General Health Questionnaire (GHQ-12) has been widely used in many settings and languages, and includes positively and negatively worded items. Our aim was to test the hypothesis that the GHQ-12 assesses both positive and negative mental health and that these domains are independent of one another.

Method. Exploratory (EFA) and confirmatory (CFA) factor analyses were conducted using data from the British Household Panel Survey (BHPS) and the Health Survey for England (HSE). Regression models were used to assess whether associations with individual and household characteristics varied across positive and negative mental health dimensions. We also explored higher-level variance in these measures, between electoral wards.

Results. We found a consistent, replicable factor structure in both datasets. EFA results indicated a two-factor solution, and CFA demonstrated that this was superior to a one-factor model. These factors correspond to ‘symptoms of mental disorder’ and ‘positive mental health’. Further analyses demonstrated independence of these factors in associations with age, gender, employment status, poor housing and household composition. Statistically significant ward-level variance was found for symptoms of mental disorder but not positive mental health.

Conclusions. The GHQ-12 measures both positive and negative aspects of mental health, and although correlated, these dimensions have some independence. The GHQ-12 could be used to measure positive mental health in population-based research.

INTRODUCTION

Well-being, and not just the absence of undesirable states such as depression or anxiety, may have consequences for physical health, cognition, relationships and even survival (Danner et al. 2001; Huppert & Whittington, 2003; Keyes, 2005). However, evidence remains sparse (Ostir et al. 2000; Strandberg et al. 2006). This is partly due to the absence of appropriate population-based measures, which in turn reflects the lack of consensus about the most salient underlying constructs or dimensions (Ryff & Singer, 1998; Ryan & Deci, 2001; Huppert & Whittington, 2003). There are two broad (and complementary) traditions in conceptualizing well-being: the ‘hedonic’ and ‘eudaimonic’ approaches. The hedonic approach emphasizes happiness (pleasant affect, life satisfaction) whereas the eudaimonic approach (with origins in Platonic philosophy) emphasizes optimal psychological and social functioning (or ‘flourishing’).

The 12-item General Health Questionnaire (GHQ-12), a brief self-report measure, has excellent psychometric properties as a screening instrument for psychiatric disorders in non-clinical settings (Goldberg & Williams, 1988). It
has been used extensively in epidemiological research, with scores taken to indicate the severity of symptoms of the most common mental disorders, anxiety and depression (Weich & Lewis, 1998; Weich et al. 2001, 2003). Factor analyses have reported one-, two- and three-factor models (Campbell et al. 2003). Many two- and three-factor models describe a pathological dimension (for example ‘anxiety-depression’ or ‘loss of confidence’; Shevlin & Adamson, 2005). Lewis (1992), examining the 30-item GHQ, noted that positive and negative GHQ items had different patterns of factor loading. He found a first (unrotated) general factor on which all items were loaded and a second, bipolar factor with positive and negative mental health at opposite ends. Huppert & Whittington (2003) and others have also suggested that positive mental health and symptoms of mental disorder are independent of one another, to a degree (Watson et al. 1988; Keyes, 2005).

Our aim was to test the hypothesis that the GHQ-12 assesses both positive and negative mental health, and that these domains are independent of one another. We set out to achieve this using two large, nationally representative samples of the British population.

**METHOD**

**Data sources and measures**

Data were taken from the first wave of the British Household Panel Survey (BHPS), recruited in 1991, and the combined 1991 and 1992 samples of the Health Survey for England (HSE). The BHPS is an annual survey of each adult member of a representative sample of private households in England, Wales and Scotland (Taylor et al. 2005). The HSE is a series of annual surveys about the health of people in England (OPCS, 1994). Like the BHPS, the HSE samples households from the Postcode Address File (PAF) using a stratified sampling frame. In both surveys, interviews were sought with all adult members of sampled households. We restricted both samples to those aged 16–74.

The 12-item GHQ-12 comprises six ‘positive’ and six ‘negative’ items concerning the past few weeks. Positive items included ‘Have you recently felt capable of making decisions about things?’, while negative items included ‘Have you recently felt constantly under strain?’ Items were classified in this way according to wording, with positively worded items having responses ‘Better than usual’, ‘Same as usual’, ‘Less than usual’ and ‘Much less than usual’. Responses to negatively worded items are ‘Not at all’, ‘No more than usual’, ‘Rather more than usual’ and ‘Much more than usual’. We regarded questions 1, 3, 4, 7, 8 and 12 as positively worded items. The remainder are negatively worded. Responses were coded using an unweighted four-point Likert scale (0, 1, 2, 3). Positively worded items were later rescored so that a high score was indicative of endorsement of these items (e.g. ‘better than usual’). Higher scores on negative items indicate greater distress and/or difficulty.

**Individual- and household-level risk factors**

Age, gender, marital status, ethnicity, education, employment status, financial strain and number of current physical health problems were included as individual-level characteristics. Household-level characteristics were: social class of the head of household, structural housing problems (any major problem, or two or more minor problems such as damp, condensation, leaking roof or rot in wood), low income (household income below half the sample median), household access to a car, housing tenure, overcrowding (more than two household members per bedroom) and household composition.

**Statistical analyses**

Exploratory factor analysis (EFA) was carried out to assess whether the GHQ-12 has one or two underlying factors. We used the principal component analysis method to extract factors. The number of factors was determined by the examination of scree plots and the size of eigenvalues. An orthogonal (varimax) rotation was then made to achieve a more readily interpretable factor structure. We chose 0.4 as a cutoff for size of loading to be interpreted. Factor analysis was undertaken using Stata (Stata Corporation, 2004).

We verified the factor structure of the GHQ-12 by creating several split samples from the study dataset (with replacement) (Diamantopoulos & Siguaw, 2000). These randomly drawn datasets were: (1) a sample of 500 observations including 250 women and 250 men; (2) and (3) samples of
500 observations for each sex, separately; (4) a sample of 500 observations from participants aged 40 years and over; and (5) a sample of 500 observations from those below 40 years of age. We first conducted separate EFAs and compared results with those derived from the original sample. We then conducted confirmatory factor analysis (CFA) in each sample. CFA is a modelling technique to investigate whether a particular factor structure is consistent with the correlations or covariance of a set of observed variables (Jacob et al. 1998). Hence, the hypothesized factor structure has to be specified a priori. The AMOS Graphics (AMOS 6, AMOS Development Corp. SPSS Inc., Chicago, IL, USA) was used. Fit indices of different models were compared. The $\chi^2$ statistic assesses an overall model fit in covariance structure models, but it is sensitive to sample size. Using a sufficiently large sample tends to yield a significant value of $\chi^2$ (Jöreskog, 1969; Bentler & Bonett, 1980), which indicates lack of satisfactory fit. Hence, the $\chi^2$ statistic has limited value (Kaplan, 1990) and was not used. Instead, we used the goodness-of-fit index (GFI), the adjusted GFI (AGFI), the root-mean-square error of approximation (RMSEA) values, and the expected cross-validation index (ECVI). GFI is based on the ratio of the sum of the squared differences between the observed and reproduced matrices to the observed variances and should be $\geq 0.90$. AGFI is a variant of GFI that adjusts GFI for degrees of freedom. RMSEA is also known as the discrepancy (between the observed and produced covariance matrices) per degree of freedom. ECVI is used to identify whether a model is likely to cross-validate samples of the same size from the same population (Diamantopoulos & Siguaw, 2000); lower ECVI indicates better fit.

Because extracted factors are based on latent variables, loadings were used to derive factor scores. These factor scores are estimates of values for unobserved random variables (Johnson & Wichern, 1998). Factor scores are obtained using a regression approach, producing the highest correlations between factors and factor scores. Each factor score had a standardized normal distribution with zero mean and unit variance.

Finally, we tested the hypothesis that scores for ‘negative’ and ‘positive’ GHQ-12 factors would show different patterns of association with demographic, socio-economic and contextual risk factors. These analyses were conducted in the full BHPS data set. To reflect the hierarchical structure of these data (individuals nested within household; households within regions, etc.) and to avoid underestimating standard errors, we used multilevel analysis using MLwiN (Snijders & Bosker, 1999; Rasbash et al. 2004). We specified three levels, comprising individuals, households and electoral wards (mean population 5222, s.d. = 899; the smallest geographical area at which BHPS data were available). Random intercept models were fitted. A null model was first estimated and individual and household’s risk factors were added subsequently. Parameters were estimated using Iterative Generalized Least Square (IGLS). Wald tests on both fixed and random parameters were performed. We also reported the 95% interval estimates for each random parameter based on normality assumption and the MCMC estimation. The MCMC method is a Bayesian formulation that puts priors on parameters and generates a correlated chain of sample draws from the posterior distribution of the parameters. When comparing two non-nested models, for example, the model with age being used as a continuous variable and the model with age being used as a categorical variable, we used the Akaike Information Criterion (AIC). The AIC is calculated based on the formula $l + 2p$, where $l$ is the value of $-2\log$ likelihood for each model, $p$ is the number of parameters fitted in the model. The model with the smallest AIC is deemed to fit best.

RESULTS

The BHPS sample comprised 9204 participants, of whom 8978 (97.5%) completed the GHQ. Among these, mean age was 41.7 years (s.d. = 16.0) and 52.7% ($n = 4735$) were female. The HSE sample comprised 6451 participants with a completed GHQ, of whom 52.9% ($n = 3412$) were women. Mean age of the HSE sample was 43.0 years (s.d. = 16.4). The sample characteristics are shown in Table 1.

EFAs produced almost identical results in the two samples (Table 2). Unrotated factor analysis revealed a first factor on which all negatively worded items loaded at above $+0.6$. The first
two unrotated factors explained 40.3% and 10.9% of the variance in GHQ-12 score in the BHPS sample, and 42.1% and 11.6% of variance in the HSE sample. Inspection of eigenvalues and scree plots indicated two-factor solutions in both samples. Factor loadings after rotation are shown in Table 2.

Table 2 shows that all negatively worded items had a high loading on the first rotated factor, with a smallest value of 0.54. For positively worded items, negative loading values on factor 1 were found, but were much lower than the loadings of negatively worded items. Loadings for item 1 (‘able to concentrate on whatever you are doing’), item 7 (‘able to enjoy your normal day-to-day activities’) and item 12 (‘feeling reasonably happy, all things considered?’) on the first factor were all \( \leq -0.40 \), with the exception of item 7 in the HSE sample (where loading on factor 1 was \(-0.38\)). The first factor therefore appears to measure symptoms of mental disorder, incorporating negative affect (feeling unhappy or depressed; worthlessness, loss of confidence, unable to enjoy normal activities), anxiety (feeling under strain, losing sleep over worry), and impaired mental functioning (poor concentration).

There was also a clear pattern to loadings on the second factor: loadings for all positively worded items had a positive sign, and all negatively worded items had a negative sign. Correlations between factor 2 and some positively worded items were very high: the loadings of items 3, 4 and 8 were above 0.6. By contrast, all negatively worded items had much lower loadings, and all were below 0.4. We judged factor 2 to represent positive mental health, incorporating positive mental functioning (good concentration; playing a useful part in things; facing up to problems; making decisions) and positive affect (feeling happy; enjoying normal daily activities).

Table 3 reports measures of fit for CFA models. The findings from both samples were again highly similar, and we have therefore only reported results from the BHPS sample (HSE results available from the authors on request). Model I has a one-factor structure and assumes that the GHQ-12 is unidimensional. Model II postulates a two-factor solution in which each negatively worded item is specified to load only on the first factor, while each positively worded item depends only on the second factor. The two factors are constrained to have zero covariance. The factor structure in model III is the same as model II, except that the covariance between two factors is free to be estimated. In keeping with the findings of our EFA, model IV was specified in which all six negatively worded items only loaded on the first factor. Among positively worded items, items 1, 7 and 12 were allowed to load on both first and second factors, as EFA showed that the correlation between these three items and the first factor was >0.4. Likewise, covariance between the two factors was not constrained in model IV. Model IV was superior on all four fit statistics for all samples (Table 3). The values of RMSEA for both
models I and II exceeded 0.1, the point at which models should not be accepted. Comparison of GFI across the models demonstrates that the proportion of the sample matrix predicted by the reproduced matrix for models III and IV was greater than 0.90. Although there were marginal differences in the values of GFI between models III and IV, ECVI shows that model IV has smaller values, and therefore had a slightly better fit. Where variance between the two factors was not constrained, correlation coefficients between these were slightly larger in model III (range 0.56 to 0.83) than model IV (0.56 to 0.73) across the subsamples specified in Table 3.

Multilevel analyses

In multilevel null models, individual-level variance was similar for scores on symptoms of mental disorder (factor 1) (0.96, s.e. = 0.02, p < 0.001) and positive mental health (factor 2) (0.95, s.e. = 0.02, p < 0.001). Approximately 4% of variance in these scores (0.04, s.e. = 0.02, p = 0.007) was found at household level for symptoms and 5% (0.05, s.e. = 0.02, p = 0.002) for positive mental health, which was statistically significant in both cases. A different pattern was observed for variances at ward level: the estimated variance in scores between wards was statistically significant for the symptoms factor (0.013, s.e. = 0.005, p = 0.003) but not for positive mental health (0.004, s.e. = 0.004, p = 0.24).

Table 4 shows associations between individual and household characteristics and symptoms (factor 1) and positive mental health (factor 2) scores. Different patterns of association were observed, particularly with age (Fig. 1), gender, ethnicity, employment, financial strain, poor housing and household composition. Female gender was associated with a higher symptoms (F1) score to a statistically significant degree, but not with lower positive mental health (F2). The same was true for non-white ethnicity, moderate financial strain, housing problems, living alone and being a single parent. Even where the association with positive mental health score mirrored that with symptoms, the effect size of the latter was substantially larger. Being retired was associated with significantly lower positive mental health and lower symptoms scores.

DISCUSSION

We found a consistent, replicable factor structure to the GHQ-12, using EFA and CFA in two large, nationally representative datasets.
Politi et al. studies have reported one- (Lewis, 1992), two-
2003; Shevlin & Adamson, 2005). Previous
studies (including Graetz, 1991; Politi et al.
2004; Shevlin & Adamson, 2005). Some of
these studies typically find inter-factor corre-
lations of $\geq 0.6$ (Cheung, 2002; Kalliath
et al. 2004; Shevlin & Adamson, 2005).

Campbell et al. (2003) used CFA to compare
seven models using data from 500 primary
care attendees in Tasmania and original in-
vestigators’ scoring methods. No model
achieved ‘good’ or ‘better than good’ fit, but all
fitted equally well. When Likert scoring was
used, the best fit was achieved by a three-factor
model in which many items loaded on two fac-
tors. These factors were described as ‘dys-
phoria’ (or ‘anxiety and depression’), ‘social
functioning’ and ‘loss of confidence’. A similar
factor structure was reported by Shevlin and
Adamson (2005) using data from a general
population survey in Northern Ireland. In the
model preferred by Shevlin and Adamson
(2005), ‘difficulty with concentration’ (item 1),
‘(not) enjoying normal activities’ (item 7) and
‘(not) feeling reasonably happy’ (item 12)
all have high positive loadings on the second
factor, entitled (perhaps erroneously) ‘social
dysfunction’. These are the same items that had
high loadings on both of the factors identified
in the present study, and were allowed to load
on both in our CFAs. Our factor descriptions
core the results of our EFAs.

Although all studies have shown that the first,
general factor explains over 40% of the variance
in GHQ-12 scores, findings about the precise
structure of this measure differ (Campbell et al.
2003; Shevlin & Adamson, 2005). Previous
studies have reported one- (Lewis, 1992), two-
(Politi et al. 1994; Kalliath et al. 2004) and
three-factor solutions (Cheung, 2002; Campbell
et al. 2003; Shevlin & Adamson, 2005). Some of
this apparent inconsistency may be methodo-
logical in origin, including differences in setting,
sample size and composition, weighting of item
scores and methods of analysis. Studies using
non-Likert scoring and/or EFA are prone to
divergent findings. The largest number of stud-
ies (including Graetz, 1991; Politi et al. 1994;
these items were loaded only on the second factor. We suggest that these two factors are best described as symptoms of mental disorder (encompassing negative affect, anxiety and impaired mental functioning) and positive mental health (covering positive mental functioning and positive affect).

Although the utility of correlated GHQ factors as subscales is potentially limited (Shevlin & Adamson, 2005), our analyses show that the two factors studied here differ from one another to a significant degree in their associations with socio-economic and demographic variables. Like Huppert and Whittington (2003), we found that associations with age were different for positive mental health and symptoms of mental disorder. Positive mental health showed a consistent decline with age for men and women, while the association with symptoms of mental disorder defied simple description. Unemployment was associated with positive mental health to a greater extent than with symptoms of mental disorder, also in keeping with previous findings (Huppert & Whittington, 2003). Positive mental health was most strongly associated with younger age, being in work, being single, not having any financial strain and having no or few physical health problems. A different picture emerged when considering those with fewer symptoms of mental disorder, who were more likely to be male, older, white, neither a single parent nor living alone, with fewer physical health problems and not experiencing financial strain. This supports the view that ‘symptoms’ and ‘positive mental health’ scores are not merely mirror images of each other.

This is the first study of which we are aware to estimate variance in positive mental health at distinct levels within a multilevel model. Around 95% of variance in both positive mental health and symptoms of mental disorder occurred at the individual level, with a further 4% occurring in:

Table 4. Associations between individual- and household-level characteristics and symptoms of mental disorder (F1) and positive mental health (F2) scores (BHPS data)

<table>
<thead>
<tr>
<th></th>
<th>Symptoms (F1)</th>
<th>Positive mental health (F2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>B (s.e.)</td>
</tr>
<tr>
<td><strong>Age (v. &gt;55 years)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16–35 years</td>
<td>3683</td>
<td>0.20 (0.04)</td>
</tr>
<tr>
<td>36–55 years</td>
<td>3218</td>
<td>0.24 (0.04)</td>
</tr>
<tr>
<td><strong>Females</strong></td>
<td>4735</td>
<td>0.18 (0.02)</td>
</tr>
<tr>
<td><strong>Non-white ethnicity</strong></td>
<td>350</td>
<td>0.12 (0.05)</td>
</tr>
<tr>
<td><strong>Marital status (v. married/cohabited)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Divorced/separated/widow</td>
<td>1015</td>
<td>0.07 (0.04)</td>
</tr>
<tr>
<td>Single (never married)</td>
<td>1870</td>
<td>−0.07 (0.03)</td>
</tr>
<tr>
<td>No educational qualifications</td>
<td>2690</td>
<td>−0.05 (0.02)</td>
</tr>
<tr>
<td><strong>Employment status (v. employed)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unemployed</td>
<td>521</td>
<td>0.06 (0.05)</td>
</tr>
<tr>
<td>Retired</td>
<td>1148</td>
<td>−0.18 (0.04)</td>
</tr>
<tr>
<td>Economically inactive</td>
<td>2837</td>
<td>0.06 (0.03)</td>
</tr>
<tr>
<td><strong>Financial strain (v. comfortable)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doing all right</td>
<td>2546</td>
<td>0.08 (0.03)</td>
</tr>
<tr>
<td>Just about getting by</td>
<td>2777</td>
<td>0.35 (0.03)</td>
</tr>
<tr>
<td>Finding it difficult</td>
<td>1237</td>
<td>0.76 (0.04)</td>
</tr>
<tr>
<td><strong>Physical health problems (v. 0)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2837</td>
<td>0.20 (0.02)</td>
</tr>
<tr>
<td>2</td>
<td>1234</td>
<td>0.27 (0.03)</td>
</tr>
<tr>
<td>3</td>
<td>463</td>
<td>0.47 (0.05)</td>
</tr>
<tr>
<td>4 or more</td>
<td>199</td>
<td>0.69 (0.07)</td>
</tr>
<tr>
<td>Low household income</td>
<td>1707</td>
<td>0.05 (0.04)</td>
</tr>
<tr>
<td>Poor housing</td>
<td>2430</td>
<td>0.12 (0.02)</td>
</tr>
<tr>
<td>No car access</td>
<td>531</td>
<td>−0.05 (0.03)</td>
</tr>
<tr>
<td>Living alone</td>
<td>973</td>
<td>0.15 (0.05)</td>
</tr>
<tr>
<td>Single parent</td>
<td>344</td>
<td>0.21 (0.06)</td>
</tr>
</tbody>
</table>

BHPS, British Household Panel Survey; s.e., standard error.
between households. Previous findings based on 1991 BHPS data estimated that 0.9% (95% CI 0.01–1.8, \( p = 0.05 \)) of variance in total GHQ-12 score occurred between wards (Weich et al. 2003). We found statistically significant between-ward variance in symptoms score (1.3% of total variance) but not in positive mental health score (0.4% of variance).

**Strengths and weaknesses**

An important strength was use of both BHPS and HSE datasets: large, representative population-based samples with high response rates. These were large enough to demonstrate consistent findings in CFAs across subsamples.

A further strength was scoring the GHQ-12 using an unweighted Likert scale. We avoided *a priori* assumptions about the relationship between individual responses and underlying GHQ-12 structure (Huppert & Whittington, 2003). In keeping with Huppert and Whittington (2003), we reversed the scoring of positively worded items such that ‘more than usual’ was scored more highly than ‘less than usual’. By contrast, traditional scoring rates the absence of positive feelings as equivalent to symptoms of mental disorder. We repeated our analyses using ‘original’ GHQ-12 scores; that is, positively worded items scored 0, 1, 2, 3 (from ‘more than usual’ to ‘less than usual’). EFAs indicated that these remained identical numerically to those shown in Table 1, but all became positive in sign. We found no evidence of any substantial effect on results shown in Tables 3 or 4 or Fig. 1. Direction of association with the ‘new’ F2 (positive mental health) score was reversed, such that higher scores reflected a reduction in positive mental health (because ‘less than usual’ attracted a higher score than ‘more than usual’ on items such as ‘able to enjoy normal daily activities’). Our findings were not due to artefact arising from the chosen scoring method. Finally, standard EFA implicitly assumes that the observed variables represent measurements that are, at least approximately (like the GHQ), on an interval scale (Everitt, 1984).

One weakness lies in the GHQ-12 itself. All self-report questionnaires are prone to method variance, namely the tendency for people to respond the same way to similarly worded items. This may contribute to the aggregation of responses to positively and negatively worded items. Although statistical methods exist for distinguishing between this and ‘true’ variance due to latent constructs, these methods are not applicable where there are only one or two construct factors. While a one-factor solution with positive and negative methods factors might have worked equally well (or better), it was not possible to test this empirically.

We have not formally established the concurrent validity of the proposed GHQ positive mental health scale. This would require testing whether the ‘positive’ items on the GHQ-12 do indeed correlate with items of positive mental health or positive emotionality on other personality scales, which was beyond the scope of the present study. Kalliath et al. (2004) reported similar positive and negative factors (although restricted to four items each) and found that scores on both were correlated to a statistically significant, and similar, degree with measures of employment and family well-being.
CONCLUSIONS

The best description of the factor structure of the GHQ-12 is given by a variant of a two-factor solution, corresponding to positive mental health and symptoms of mental disorder. A more precise specification reveals that the former encompasses both positive affect and positive mental functioning. This dimension has been shown to exhibit substantial, clinically meaningful independence from measures of symptoms (the more commonly measured, pathological, ‘negative mental health’ dimension). The present study provides support for the view that valid measures of positive mental health should include items that assess both the hedonic and eudaimonic domains – positive affect as well as positive social and mental functioning. While a two-dimensional model may not offer much of an advantage over a one-factor model when screening for psychiatric disorders, the former approach may come into its own in studying the determinants and consequences of well-being.

DECLARATION OF INTEREST

None.

REFERENCES


