1 **Understanding the population structure of the GHQ-12: evidence for**
2 **multidimensionality using Bayesian and Exploratory Structural Equation Modelling**
3 **from a large-scale UK population survey.**

4

5 Gareth J Griffith^{a†}, Kelvyn Jones^a

6 ^a School of Geographical Sciences, University Road, University of Bristol, BS8 1SS, United Kingdom

7

8 [†] Corresponding author: Gareth Griffith. Email: g.griffith@bristol.ac.uk ,

9 Address: as above, Tel: (0)+44 7749 338138

10

11

12

13 **Abstract**

14 Mental health and its complexity, measurement and social determinants are increasingly important
 15 avenues of research for social scientists. Quantitative social science commonly investigates mental
 16 health as captured by population screening metrics. One of the most common of these metrics is the
 17 12-Item General Health Questionnaire (GHQ-12). Despite its canonical use as an outcome of interest
 18 in social science, the traditional use of the summed scores of summed questionnaires carries
 19 empirical and substantive assumptions which are often not fully considered or justified in the
 20 research. We outline the implications of these assumptions and the restrictions imposed by
 21 traditional modelling techniques and advocate for a more nuanced approach to population mental
 22 health inference. We use novel Exploratory Structural Equation Modelling (ESEM) on a large,
 23 representative UK sample taken from the first wave of the Understanding Society Survey, totalling
 24 40,452 respondents. We use this to exemplify the potential of traditional, restrictive assumptions to
 25 bias conclusions and policy recommendations. ESEM analysis identifies a 4-factor structure for the
 26 GHQ-12, including a newly proposed “Emotional Coping” dimension. This structure is then tested
 27 against leading proposed factor structures from the literature and is demonstrated to perform
 28 better across all metrics, under both Maximum Likelihood and Bayesian estimation. Moreover, the
 29 proposed factors are more substantively dissimilar than those retrieved from previous literature.
 30 The results highlight the inferential limitations of using simple summed scores for mental health
 31 measurement. Use of the highlighted methods in combination with population studies offers
 32 quantitative social scientists the opportunity to explore predictors and patterns of underlying
 33 processes of population mental health outcomes, explicitly addressing the complexity and
 34 measurement error inherent to mental health analysis.

35

36 Key words (7) : UK, Mental Health, Depression, Structural Equation Modelling, Population Health,
 37 GHQ-12.

38

391 INTRODUCTION

40 Mental health measurement, specifically measurement of mood disorders, is notoriously
 41 difficult. This difficulty is largely due to the challenges of quantifying an individual's mental
 42 health status robustly and the presence of ambiguity around tangible thresholds. As a result it
 43 is likely that there is substantial under-diagnosis of mental health disorders with as many as
 44 74-85% of depressed individuals never receiving diagnoses of depression (Lecrubier, 2007;
 45 Verheij, 1996). It follows that the most under-reported group are individuals with 'milder'
 46 cases of depression (Garrard et al., 1998). As clinical outcomes under-report psychological
 47 morbidity, quantitative social scientists commonly address this by estimating psychological
 48 morbidity from questionnaires distributed to a sample of the population.

49 The 12-Item General Health Questionnaire (GHQ-12) is one of the most widely used
 50 responses in quantitative social science and epidemiology for the analysis of mental health
 51 trends. It constitutes 12 questions, each with 4 Likert response categories, which when
 52 answered are conventionally summed to give a single overall score lying on a notionally
 53 meaningful singular dimension (Goldberg and Hillier, 1979). The popularity of the GHQ-12
 54 is mostly due to its ease of use, and capacity to reliably reproduce "remarkably robust" results
 55 contrasted with longer initial versions (Goldberg et al., 1997). Initially focusing on diagnostic
 56 purposes for specifically at-risk individuals, the GHQ-12 has since been translated and
 57 validated across multiple languages and countries as a screening tool for depression and
 58 depressive symptoms across populations (Creed and Evans, 2002; Hankins, 2008a; Pevalin,
 59 2000; Smith et al., 2010). The UK is no exception to this, where the GHQ-12 has become the
 60 canonical mental health measure for UK-based population studies (Propper et al., 2005;
 61 Thomson and Katikireddi, 2018; Weich et al., 2003), largely due to its inclusion as principal
 62 mental health outcome for a series of popular large-scale surveys including the UK

63 Household Longitudinal Survey (McFall, 2011) and the Health Survey for England (Mindell
64 et al., 2012).

65 Despite being extensively validated using other leading mental health metrics, as well as
66 longer versions of the same GHQ metric, there is still considerable debate as to what the
67 GHQ-12 is truly capturing (Werneke et al., 2000; Ye, 2009). Due to its repurposed nature,
68 different disciplines treat the GHQ-12 differently, seeking either to identify cases of ill-health
69 (Goldberg and Bridges, 1987), or understand population trends across mental health spectra
70 (Hu et al., 2007; Weich et al., 2003). This has led to numerous multidimensional factor
71 structures being proposed in the literature, with little consensus on whether more complex
72 structures are truly adding value or if they result simply from over-interpretation of
73 substantively meaningless stochastic variation (Aguado et al., 2012; Hankins, 2008a).

74 The critical importance of understanding composite and complex measures such as the GHQ-
75 12 has been clearly evidenced in recent genetic and psychological research. For instance,
76 more strictly defined cases of depression produce higher estimates of genetic heritability for
77 major depressive disorder (Cai et al., 2018); and mental illness and wellbeing, often used
78 interchangeably in policy, can have widely differing covariates (Patalay and Fitzsimons,
79 2016; Westerhof and Keyes, 2010). Greater nuance in understanding has also been explicitly
80 advocated by the UK Chief Medical Officer, who called for greater understanding of both
81 positive and negative components of mental health (Davies, 2013, 2018).

82 This paper uses novel Exploratory Structural Equation Modelling techniques to estimate
83 underpinning processes governing the GHQ-12 in the UK. This allows the relaxation of
84 previously necessary methodological constraints which have been demonstrated to bias
85 results towards simpler interpretations (Marsh et al., 2014). The factor structure provided by
86 this procedure is subsequently evaluated against leading interpretations of GHQ-12

dimensionality. Moreover, all models are estimated using both Maximum Likelihood (ML) and Bayesian procedures allowing for more flexible estimation of non-normal values, giving less biased results for potentially skewed factor variances (Muthén and Asparouhov, 2010)

The rest of the paper is organised as follows. Firstly, a brief overview of the methodological assumptions underpinning traditional factor analysis methods is provided. We then critically review previous structures found in the literature for the GHQ-12, identifying common dimensions found in studies across sub-populations. We then consider the large-scale data to be analysed from Understanding Society and briefly outline a recently developed analytical approach, ESEM, which allows a less constrained approach to instrument decomposition than traditional factor analysis, with the potential to reveal a more nuanced structure. This analytical approach is then applied to identify an optimal factor structure for the study population, which is then compared with structures proposed and validated by previous studies. An extended discussion section concludes by exploring the implications and limitations of this newly found structure, with recommendations for best practice for quantitative social scientists in generating meaningful insight from complex mental health outcomes.

1.1 *Understanding Mental Health Measurement*

The desire for population screening metrics can be seen as symptomatic of a wider cultural shift in mental health perception away from an “absence of illness” perspective. Where early studies of the ecology of mental illness focused on diagnoses of psychosis (e.g. Faris and Dunham, 1939), evidencing change via reduction in illness, contemporary approaches advocate a more holistic approach aiming to measure improvements in “wellbeing” (World Health Organization, 2013). This necessarily involves the conceptualisation and characterisation of a mental health “spectrum”, beyond binary caseness. Whilst complexity is

increasingly recognised as an intrinsic part of mental health measures (Gnambs and Staufenbiel, 2018; Hu et al., 2007), quantitative social science commonly takes the reductive view of reducing questionnaire responses to binary “cases” or as unidimensional constructs without adequate consideration of the methodological implications of doing so.

The underpinning theoretical assumption of using summed GHQ-12 scores is that mental health lies on a single spectrum, and that all variation in the outcome of interest occurs along that spectrum. This is referred to as unidimensionality and is particularly pertinent in literature which recommends or posits thresholds for the GHQ-12 above which an individual is considered at risk, or an “ill case” (Baksheev et al., 2011; Goldberg et al., 1998; Tait et al., 2003). In this literature, for caseness to be meaningful at the population level it is particularly critical that a unit increase can be assumed to be capturing the same change in mental status across individuals, an assumption that necessarily posits substantive equivalence in unit response change *between* items (unidimensionality), and *within* items (additivity). Ultimately all research treating the GHQ-12 as unidimensional, continuous or binary, necessarily requires the belief that a unit-increase in summed GHQ-12 score implies the same change wherever it occurs across the metric (Brodersen et al., 2007; Marsh and Bailey, 1991).

Moving beyond internal validity, the context-sensitivity of the interpretation of the GHQ-12 has been evident for some time (Werneke et al., 2000). Goldberg et al. (1998) noted the context-specific nature of thresholds in the GHQ-12, documenting the clear differences in specificity and discriminant capacity between-items across-countries. Despite this early acknowledgement, the problematic issue has been largely disregarded in subsequent literature due to necessary impositions of overly-simplistic analyses and validation measures.

Numerous studies have attempted to characterise the internal consistency of the GHQ-12 using relatively simple tests of Cronbach’s Alpha (Cronbach, 1951) or by conducting

Confirmatory Factor Analysis (CFA) (Jöreskog, 1969). Of the two, CFA is considered the superior measure, providing robust evaluations of capacity for model replication (Dunn et al., 2014). However, it still requires a refinement of dimensionality to “simple structure”, implying zero cross-loadings across items (Asparouhov and Muthén, 2009a; Marsh et al., 2014). This means that each constituent item can only be empirically related to one underpinning construct, which has in turn been argued to lead to a reliance on overfitting models via model modification indices (MacCallum et al., 1992), this renders the ostensibly *confirmatory* analysis theoretically *exploratory* (Fabrigar et al., 1999; Schmitt, 2011). The issues surrounding this approach to social science metrics are not solely empirical, as there are also substantive concerns when using traditional CFA approaches (Conway and Huffcutt, 2003). The most pressing of these concerns is the imposition of zero cross-loadings, which are particularly unrealistic for measures such as the GHQ-12, as “nonzero cross-loadings are inherent in psychological measurement” (Marsh *et al.*, 2014, pp.88).

1.2 Proposed Factor Structures

Whilst longer versions of the GHQ are commonly accepted to be multidimensional (Graetz, 1991; Martin, 1999), consensus is less forthcoming for the GHQ-12. Of the variety of proposed interpretive structures suggested for the GHQ-12, there are three common interpretations. The most simplistic involves interpreting the GHQ-12 score as a unidimensional construct, taking the summed scores as a response, occasionally with an adjustment for positive and negative phrasing. This approach is backed up by a large body of research, which uses CFA to conclude that the unidimensional interpretation of the GHQ-12 is the most compelling (e.g. Aguado et al., 2012; French and Tait, 2004; Winefield et al., 1989). Despite this research concluding in support of unidimensional interpretation there is another consideration of importance here.

Studies commonly cite high correlations between modelled multidimensional factors as justification for simplifying to unidimensional interpretation (e.g. Gao *et al.*, 2004; Gouveia *et al.*, 2010; Padrón *et al.*, 2012; Fernandes and Vasconcelos-Raposo, 2013; Romppel *et al.*, 2013). Recent work using simulated data has called this justification into question, demonstrating this imposition of simple structure to artificially inflate correlations between modelled factors (Asparouhov and Muthén, 2009a; Marsh *et al.*, 2014). Thus whilst it is not uncommon for reported correlations to be greater than 0.9 (e.g. Aguado *et al.*, 2012; Campbell and Knowles, 2007) or 0.95 (e.g. Sweeting *et al.*, 2009; Wang and Lin, 2011), taking these high correlations as justification for unidimensionality risks a self-fulfilling prophecy of simplicity begetting simplicity.

Whilst many studies call for unidimensional interpretation there are also numerous proposed multidimensional GHQ-12 structures. Several two-factor solutions have been proposed and validated using CFA techniques. These two factors most commonly involve a “Depression/Anxiety” construct and a “Social Dysfunction” construct (Andrich and Schoubroeck, 1989) (given by GHQ-12 items 2, 5, 6, 9, 10 and 11, and 1, 3, 4, 7, 8 and 12 respectively - see supplementary materials for full item list with response coding). “Depression/Anxiety” relates to the emotional component of psychological distress, whereas “Social Dysfunction” relates to the social functioning component of the individual experiencing the distress. This structure, albeit with different labelling, has at various points been identified as the best fit in data from the UK (Smith *et al.* 2010), New Zealand (Kalliath *et al.*, 2004), Brazil (Gouveia *et al.*, 2010), Japan (Suzuki *et al.*, 2011), Germany (Schmitz *et al.*, 1999), Italy (Politi *et al.*, 1994) and Turkey (Kiliç *et al.*, 1997). Systematic reviews and meta analyses also consistently identify these two factors most commonly in both two- and three-factor solutions (Gnambs and Staufienbiel, 2018; Picardi *et al.*, 2001; Werneke *et al.*, 2000). It is important to note that these groupings also align with the positive or negative

phrasing of the constituent items. “Social Dysfunction” items are all positively worded and “Depression/Anxiety” items are all negatively worded, and as such there has been debate as to whether this structure simply reflects differences in phrasing (Hankins, 2008b).

Increasing the multidimensional complexity brings consideration of three-factor solutions, which most commonly identify “Social Dysfunction” and “Depression/Anxiety” constructs alongside a third construct referring to some variant of “Loss of Confidence”. This structure was initially identified by Worsley and Gribbin (1977) in the first factor analysis of the GHQ-12. They found the “Loss of Confidence” construct alongside “Social Performance” and “Anhedonia-Sleep Disturbance” (the latter of which approximate the Social Dysfunction and Depression/Anxiety factors respectively). For Worsley and Gribbin this third posited dimension was constructed of four items (6, 9, 10 and 11) all pertaining to feeling worthless or depressed, or losing confidence, evidence of which has been supported by subsequent work (Campbell et al., 2003; Penninkilampi-Kerola et al., 2006; Vanheule and Bogaerts, 2005). Notably, Worsley and Gribbin did not subsequently refine to a simple structure solution, as there were no other structures to test against, so this solution contains non-zero cross loadings.

A similar structure was published by Graetz in 1991. Graetz analysed a large scale (N= 8998) Australian sample, resulting in a three-factor structure comprising “Anxiety/Depression”, “Social Dysfunction” and “Loss of Confidence”. The Loss of Confidence factor is similar to that of Worsley and Gribbin – but is presented in the simple structure format of CFA, with non-zero cross-loadings eliminated. Graetz’s Loss of Confidence factor only loads on items 10 and 11, which are the only two of Worsley and Gribbin’s factor that do not also load on another factor. This 3-factor structure has been the most widely accepted multi-dimensional structure since its introduction, supported by subsequent work both in the UK (e.g. Cheung,

2002; Martin & Newell, 2005; Shevlin & Adamson, 2005) as well as elsewhere (Gnambs and Staufenbiel, 2018; Padrón et al., 2012).

Whilst several proposed structures from the literature are strongly validated, there are sampling issues with their initial design. Firstly, with the exception of the Graetz 3-factor model (Graetz, 1991), structures are commonly generated and subsequently validated using data with a small sample size and limited contextual scope. It is not uncommon for validation studies to report sample sizes under 500 (e.g. Khan et al., 2013; Martin and Newell, 2005). Moreover, studies with greater statistical power commonly draw samples from heavily context-specific populations such as primary care users (e.g. Werneke *et al.*, 2000) or high-school students (e.g. Suzuki *et al.*, 2011), and not the general population. Secondly, there is little methodological agreement on effective measures and fit criteria for recommending any given structure over another, despite efforts to establish consistent approaches (Conway and Huffcutt, 2003; Hu and Bentler, 1999; Marsh et al., 2004).

There is clear substantive interest to both social and health researchers in moving beyond understanding the patterning of observed, aggregated measures of complex phenomena such as mental health. It is of critical importance to develop a better understanding of the underpinning processes driving these measures. Empirically, this means partitioning variance in response to each item into common and unique variation across questions, thus factoring for individual variability in response (Conway and Huffcutt, 2003). There is clearly a wealth of longitudinal data available for the GHQ-12 in the UK, collected alongside rich demographic and spatial data in several panel studies. In light of newer methodological developments in the modelling of complex responses, this wealth of data needs comprehensively re-evaluating to ensure it is fully exploited. This study proposes a multidimensional interpretation of latent factors patterning GHQ-12 responses, allowing

232 insight into underpinning processes rather than manifest outcomes. It further aims to clarify
233 and make explicit the empirical and substantive benefits of doing so, in using novel
234 methodological developments to consider more complex dimensionality, in the quantitative
235 analysis of a subject as nuanced and individually heterogeneous as mental health.

236

237 **2 DATA AND METHOD**

238 **2.1 Data**

239 This study uses the first wave of the large scale Understanding Society Survey (US), a
 240 nationally representative annual panel survey of over 43,000 UK respondents from over
 241 26,000 households (McFall, 2011). Data collection for Wave 1 took place between January
 242 2009 and March 2011. The GHQ-12 was administered via face-to-face interview and
 243 completed by 40,452 respondents. For a comprehensive overview of Understanding Society
 244 Data Collection and methods see Quality Profile (Lynn and Knies, 2016). Table 1 gives the
 245 response characteristics of the survey sample. For initial exploration sample weights were
 246 taken and applied from Wave 1. Sample weights were applied to all but Confirmatory
 247 Bayesian Analyses, as weights are not supported in Bayesian estimation.

Response Sample	40452 Individuals	
Age	Mean = 45.8, S.D = 18.0	
Sex	Male 43.8%	Female 56.2%
Ethnicity (Ethnic)	White 83.1%	Black 4.6%
	Asian 9.1%	Mixed 3.3%
	Unemployed 43.5%	Professional 3.6%
Job Classification	Managerial/Technical 20.9%	Skilled Non-Manual 12.6%
	Skilled Manual 9.0%	Partly Skilled 8.2%
	Unskilled 1.9%	
Highest Educational Qualification	Higher Education 33.8%	A-Level 19.1%
	GCSE 20.9%	Other Qualification 5%
	None 21.2%	

248 *Table 1: Response Characteristics of Survey Sample completing the GHQ-12 (Understanding Society Wave 1)*

249 There is strong evidence of different response patterning across positive and negative
 250 questions as part of the GHQ-12, shown in Figure 1. It can be seen that there are strongly
 251 different modal responses for positive and negative items, giving rise to the common factor
 252 structures reflecting positive and negative items. Despite strong patterning in responses which

253 is somewhat corrected for in binary interpretation, Likert scores have been demonstrated to
 254 allow greater discrimination between different dimensional structures (Campbell and
 255 Knowles, 2007; Smith et al., 2013). It is for this reason that we use full information Likert
 256 scoring here. Items are all coded such that higher scores indicate greater distress.

Figure 1: Graph illustrating the differential frequency of response patterning of positively (1,3,4,7,8,12) and negatively (2,5,6,9,10,11) phrased items in the GHQ-12 from 2009 Wave of Understanding Society.

257

2582.2 *Methodology*

259 To explore the GHQ-12 data a novel approach termed Exploratory Structural Equation
 260 Modelling (ESEM) is deployed (Asparouhov and Muthén, 2009). ESEM represents a
 261 combination of the best elements of restrictive CFA and unstructured exploratory factor
 262 analysis (EFA). The key contribution of ESEM methodology for this research is the
 263 specification of non-zero cross-loadings on constituent items (Asparouhov and Muthén,
 264 2009). As outlined above, in a literature which commonly cites high factor correlations as
 265 justification for refining to unidimensional interpretation, it is especially important to guard
 266 against inflated estimates of factor correlations which result from imposing zero cross-
 267 loadings (Asparouhov et al., 2015; Marsh et al., 2014, 2010). Beyond the empirical, there is
 268 also a substantive argument for not refining to simple structure when analysing complex
 269 social outcomes such as psychological constructs. As multiple and inter-related underpinning
 270 processes are likely to give rise to any specific outcome most items will have multiple
 271 determinants and thus nonzero cross-loadings ought not to be viewed an aberration but a
 272 logically anticipated representation of complex, underpinning constructs (Marsh *et al.*, 2014).
 273 ESEM is carried out within the Mplus software environment (Muthén and Muthén, 1998-
 274 2018).

275 The model is estimated using a two-step procedure. Firstly, a series of EFA solutions in
 276 which each item loads on each and every of 2-5 constructs is estimated using Geomin
 277 rotation ($\epsilon = 0.01$), the optimal rotation when little is known about the true underlying
 278 structure, and when suspected variable complexity is greater than one – that is, there is an
 279 expectation that there will be cross-loadings (Browne, 2001; McDonald, 2005). Loading
 280 retention is dictated by a simple cutoff of EFA loadings <0.15 . As the need to exclude
 281 loadings purely for methodological reasons has been relaxed, the value of 0.15 is a far more

tolerant exclusion criterion than typically advocated in the literature. To test this criterion, models were rerun with all loadings present but target rotation of 0 for these small EFA loadings. This resulted in all non-target loadings having absolute values below 0.1, thus they were omitted in the confirmatory component of the model comparison.

The Likert GHQ-12 responses are specified as ordinal variables and ML estimation of confirmatory analyses was carried out using the mean and variance-adjusted weighted least squares estimator as this presents the best method for categorical or ordered data (Schmitt, 2011). This estimation specifies a probit regression equation for each item on the related factor (Muthén & Muthén, 1998-2011, pp. 62). These can be interpreted as likelihood changes in the log-odds of changing response category on the response variable, illustrating the strength of the relation between the underpinning dimension and the probability of response change in the associated item (Chen, 2007).

The models were all repeated using Bayesian estimation, allowing for less biased estimation of non-normal response variances than common ML methods (Muthén and Asparouhov, 2010). The distributions of estimates are highly likely to be non-normal due to the constrained nature of variances and correlations, with variances having a lower limit of zero, and correlations bounded by 1 and -1. Bayesian estimation does not require the ML assumption of normality of the parameter estimates, with prior variance-covariance estimates instead drawn from an inverse-Wishart distribution (Asparouhov and Muthén, 2010).

Model fit for the ML estimation was evaluated using the Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), Root Mean Square Error of Approximation (RMSEA) and the Weighted Root Mean Square Residual (WRMR). Evidence of good fit was taken from guidelines proposed by Bentler (2007) and Muthén and Muthén (2014). For the Bayesian models, fit is evaluated using Posterior Predictive Checking (PPC) for which we present

posterior predictive p-values and 95% credible intervals for each model. In Mplus PPC is an extension of the likelihood ratio statistic taken to be an indication of the model's capacity to reproduce the data and summarise the posterior distribution of the residuals (Asparouhov and Muthén, 2010). As such, the statistic is upward biased by sample size so we consider evidence of improvement in predictive capacity as reduction towards zero of the predictive credible interval (Gelman et al., 1996; Marsh et al., 2004). We also present mean absolute factor correlation values as evidence of substantive dissimilarity between modelled constructs with lower values indicating greater discriminant validity.

314

3153 RESULTS

316 Initial analysis specified a series of EFAs with between 2 and 5 factors, the results of which
 317 are given in Table 2. In the traditional ML estimation, goodness of fit statistics evidence
 318 better model improvement for each added dimension up to a five-factor solution. In the
 319 Bayesian estimation, it is more explicit that none of the specifications provide an adequate fit
 320 to reproduce the 40,452 individuals' mental health response, as shown by the consistent
 321 posterior p-values of 0.000. However, there is evidence of improvement of fit in the posterior
 322 credible interval. Predictive capacity improves with added factors up to the fourth factor;
 323 however, the five-factor solution is less robustly estimated. Reduction in the lower bound
 324 shows that a fifth factor could serve to improve predictive capacity but the increased upper
 325 bound suggests it is more probable that it will reduce the predictive capacity of the model.
 326 Therefore, it is the four-factor structure that is carried forward as it offers the most robust
 327 reproduction of observed values.

<i>EFA</i> Number of Factors	<i>Maximum Likelihood</i>				<i>Bayesian</i>		
	CFI	TLI	RMSEA	SRMR	Posterior P-Value	2.5% CI	97.5% CI
2	0.976	0.964	0.080	0.036	0.000	2668.266	2993.799
3	0.993	0.987	0.048	0.017	0.000	457.804	620.840
4	0.997	0.992	0.037	0.011	0.000	197.311	331.851
5	0.999	0.995	0.029	0.007	0.000	55.562	1073.603

328 Table.2: Bayesian and ML Fit statistics for EFA factor solutions for the GHQ-12 with 2, 3, 4 and 5 factors.

329 Having identified that the four-factor solution was the most appropriate from the estimation
 330 of the initial EFA whilst all cross-loadings are specified and estimated, the subsequent
 331 solution needed refining to a testable structure. This was implemented by re-specifying the
 332 model with all cross-loading values under 0.15 being omitted. The loadings can be
 333 interpreted as probit regression coefficients of each item on the unit-standardised response
 334 factor and are directly comparable both within and between factors.

GHQ-12 Items	F1 Lowered Self Worth	F2 Social Dysfunction	F3 Stress	F4 Emotional Coping
1. Able to Concentrate		0.659 (0.008)	0.227 (0.009)	
2. Loss of Sleep	0.632 (0.014)		0.427 (0.007)	
3. Playing a useful part		0.646 (0.007)		
4. Capable of decisions		0.907 (0.013)		-0.322 (0.017)
5. Constantly under strain	0.753 (0.018)		0.617 (0.008)	
6. Problem overcoming difficulties	0.754 (0.011)		0.330 (0.006)	
7. Enjoy day-to-day activities		0.691 (0.009)	0.200 (0.008)	0.244 (0.010)
8. Ability to face problems		0.746 (0.006)		
9. Feel unhappy/depressed	0.746 (0.009)		0.261 (0.006)	0.216 (0.009)
10. Losing confidence	0.909 (0.006)			
11. Think of self as worthless	0.887 (0.007)			
12. Feeling reasonably happy		0.577 (0.010)		0.295 (0.011)

Table 3: Standardised Factor Loadings for 4-Factor ESEM Model of the GHQ-12 using Bayesian estimation. Standard Errors in parentheses.

Table 3 gives the loadings for the four-factor ESEM solution. The four factors are labelled “Lowered Self Worth”, “Social Dysfunction”, “Stress” and “Emotional Coping”. These were chosen in accordance with existing names for previously identified factors from the literature (Graetz, 1991; Martin, 1999). The main difference from previous structures in the literature is that items 1, 2, 4, 5, 6, 7, 9 and 12 now load on multiple underlying dimensions. The loadings of Factors 1 and 2 are consistent with previous literature, and thus have been given similar names to reflect this. However, they differ empirically because they contain cross-loading items, but the broad differentiation between the positive and negatively worded items holds true. Higher individual factor scores across Factors 1 and 2 indicate higher levels of mental distress.

The third factor, here termed “Stress” is seen to load most strongly on items associated with feeling under strain and loss of sleep. It is very similar to a factor found in one of the earliest factor analyses of the GHQ-12 by Worsley and Gribbin (1977), differing only in that they found it also loaded on item 12.

The emergence of the fourth factor, termed “Emotional Coping” is a distinctive finding. It is most notable for having both positive and negative loadings. It is *negatively* associated with Item 4 – “feeling capable of making decisions”, but *positively* associated with feeling unhappy or depressed, not enjoying day-to-day activities and not feeling happy. As such individuals with high scores on Emotional Coping are those who feel negative as captured by items 7, 9 and 12, whilst feeling capable of making decisions as captured by item 4, indicating a degree of (at least perceived) perseverance in the face of the distress. The negative loading highlights an ambiguity in interpreting latent variables. That is, the interpretation is imposed by the researcher, thus an empirically identical interpretation would be the inverse, with negative loadings on Items 7, 9 and 12 and a positive loading on Item 4. This would in turn invert the poles of the underlying coping dimension, with higher scores indicating positive rather than negative outcomes. Emotional Coping is structurally most similar to the “Sleep Disturbance/Anhedonia” construct found in work by Worsley and Gribbin (1977) although there is a key difference in the negative loading of Item 4, which is clearly associated with positive functioning.

Having detailed the theoretical implications of the proposed factor structure it is necessary to understand the substantive and empirical implications of the structure. To do this we evaluate the substantive dissimilarity between factor constructs as measured by the correlations between those constructs.

<i>FACTOR</i>	F1	F2	F3	F4
---------------	----	----	----	----

<i>CORRELATIONS</i>	Lowered Self-Worth	Social Dysfunction	Stress	Emotional Coping
F1. Lowered Self Worth	1.000			
F2. Social Dysfunction	0.680	1.000		
F3. Stress	0.178	0.111	1.000	
F4. Emotional Coping	0.411	0.487	0.271	1.000

Table 4: Modelled Factor Correlations from the Four Factor ESEM Solution for the GHQ-12

370

371 Table 4 shows the modelled factor correlations. The factors with the highest correlation are
 372 Lowered Self Worth and Social Dysfunction, with a coefficient of 0.68. Although it is the
 373 highest identified in this structure, it is still low relative to that found in the existing literature.
 374 Stress is the most statistically dissimilar and therefore most substantively distinct factor,
 375 exhibiting uniformly low correlations with the other constructs.

376 It should be noted that correlation coefficients overestimate the predictive capacity of each
 377 latent variable on another. The absolute proportion of variation in one variable that could be
 378 predicted solely from knowing the other is given by the squared value of the correlation
 379 coefficients (Kish, 1954). For instance, knowing the modelled Lowered-Self-Worth scores
 380 for all individuals would only allow the prediction of 46.24% (0.4624) of the variation in
 381 Social Dysfunction scores, despite these factors having the highest modelled correlation of
 382 0.68. This is even more stark for the correlation between Stress and Social Dysfunction, with
 383 a true predictive capacity of 1.2%, leaving 98.8% of variation in one unexplained by the
 384 other.

385 Having identified a parsimonious summary of data from our model it is important to connect
 386 back to the wider literature on mental health structures. The literature has proposed many

different structures, seen in Table 5, which gives context to seven exemplar proposed structures against which we evaluate the structure proposed here.

<i>Authors</i>	<i>Date</i>	<i>N</i>	<i>Factor Structure Details</i>
Initial GHQ-12 Goldberg Formulation	1972	200	1 Factor – Baseline unidimensional model, all items specified to load on a single factor.
Hankins	2008	3705	1 Factor – Unidimensional but with error covariance specified on negatively phrased items.
Andrich & Van Schoenbroeck	1989	491	2 Factor – Split into positive and negative items, each constituting a separate dimension.
Kilic et al.	1997	1307	2 Factor – Anxiety/Depression, Social Dysfunction
Worsley and Gribbin	1977	603	3 Factor – Anhedonia-Sleep Disturbance, Social Performance, Loss of Confidence, specifies cross loadings on items 1 (F1, F2), 6 (F1, F3), 9 (F1, F3) and 12 (F1, F2)
Graetz	1991	8998	3 Factor – Anxiety/Depression, Social Dysfunction, Loss of Confidence
Sanchez-Lopez & Dresch	2008	1001	3 Factor – Successful Coping, Self-Esteem, Stress, Includes one non-zero cross-loading on Item 9 for F2 and F3.

Table 5: Seven exemplar studies of the range of factor analytical structures obtained from the GHQ-12 data. See Supplementary Materials for diagrammatic representation of structures.

Table 7 presents fit statistics for all seven models alongside the proposed ESEM solution. The mean absolute factor correlation is also presented for each specification to provide an estimate of the dissimilarity of the factors estimated in the model structure. The four-factor solution provides the best fit across every measure of fit under both traditional and Bayesian estimation.

The best performing structure outside of the ESEM solution is the original Worsley and Gribbin (1977) structure with its non-zero cross-loadings. Factor structures which address the inter-dependencies of the items via error covariance or non-zero cross-loadings perform well across all the fit statistics, which again is an argument for the adoption of a more realistically complex specification of mental health. The benefit of cross-loadings is most clearly borne out in the mean absolute factor correlation, which is far lower for the four-factor structure than other solutions.

Factor Structure	No. of Factor s	WRMR	CFI	TLI	RMSE A	Posterior predictive P	Lower Bound	Upper Bound	Mean Absolute Factor Correlation
<i>Unidimensional</i>	1	14.776	0.914	0.895	0.146	0.000	7123.436	7634.537	-
<i>Hankins</i>	1	6.808	0.981	0.967	0.082	0.000	1505.348	1764.743	-
<i>Kilic et al.</i>	2	12.826	0.934	0.918	0.129	0.000	5268.267	5709.397	0.822
<i>Andrich & van Schoubroeck</i>	2	9.018	0.967	0.959	0.091	0.000	3521.095	3892.429	0.777
<i>Sanchez-Lopez & Dresch</i>	3	8.246	0.972	0.963	0.087	0.000	2824.202	3152.636	0.772
<i>Graetz</i>	3	7.917	0.974	0.966	0.083	0.000	2459.754	2769.927	0.783
<i>Worsley & Gribbin</i>	3	5.505	0.986	0.981	0.063	0.000	1199.430	1430.393	0.665
<i>ESEM 4Fac</i>	4	3.247	0.995	0.992	0.039	0.000	376.571	526.539	0.356

Table 6: Bayesian and ML Fit Statistics for the 8 evaluated factor structure defined in Table 6.

4064 DISCUSSION

407 The results of this study show that of the factor structures tested here the four-factor structure
 408 provides the best description of GHQ-12 responses from the Understanding Society data,
 409 Wave 1. Additionally, this is evidenced not just by traditional fit statistics, but by the
 410 modelled factors being the most substantively dissimilar as evidenced by the mean absolute
 411 factor correlation. This involves the specification of two previously underexplored constructs,
 412 here termed “Stress” and “Emotional Coping”. Emotional Coping is particularly notable as
 413 the presence of a negative loading evidences the capacity for underpinning constructs to mask
 414 the presenting of psychological distress in the aggregated metric.

415 Within the wider GHQ-12 literature there is little mention of dimensions analogous to the
 416 Stress and Emotional Coping structures beyond Worsley and Gribbin’s early analysis (1977).
 417 Stress-related constructs are also proposed in a structure drawn from a Spanish population
 418 (Sánchez-López and Dresch, 2008), and notably the “thematic analysis” of GHQ-12 content
 419 by Martin (1999). These constructs would not have been discoverable using traditional CFA
 420 techniques as they consist largely of cross-loadings. The incorporation of these constructs
 421 considerably improved model fit, and the low modelled factor correlations seems to suggest
 422 that they are capturing substantively different processes. These low correlations in the results
 423 seem to support suggestions that high correlations may simply be an artefact of a restrictive
 424 modelling procedure and as such we caution refining to simpler structures based purely on
 425 apparently high factor correlations (Marsh et al., 2009).

426 The key message of this study is the capacity of large-scale datasets to contribute more
 427 comprehensive understandings of mental health outcomes in large, heterogeneous
 428 populations. Whilst this is not a new idea (Hu et al., 2007; Mukuria et al., 2014), the adoption

of less stringent exclusion criteria in model selection and the incorporation of ESEM methodologies is something that is underexplored in large-scale survey analysis. Moreover, the statistical power afforded by these large-scale surveys using decomposed metrics has the potential to offer a more comprehensive understanding of the similarities between different underpinning processes across as well as within metrics.

That said, it is important to highlight the limitations of this study. Whilst the structure proposed here provides the best fit for the data in Wave 1 of Understanding Society in the UK (McFall, 2011), it is necessary to test on a wider range of data beyond this spatiotemporally specific dataset. More research is required in order to understand fully what can be gained from the modelling of decomposed processes underpinning survey instruments. Firstly, whilst sample weights were used in the generation of the factor structure, the structure is still specific to the current dataset. Another of the key contributions of the ESEM method is the capacity for comprehensive tests of measurement invariance over subgroups within a population (Marsh et al., 2009). Whilst it has been demonstrated that this structure provides a superior fit across the full population, the degree of measurement invariance for the structure should be investigated across different geographical and demographic groups. There is considerable literature validating the overall GHQ-12 as a screening instrument over time and space (e.g. Gnambs and Staufenbiel, 2018), there is far less written on the temporal or geographical stability of its underpinning latent structure (Goldberg et al., 1997). The structure should be further validated across spatio-temporal contexts using the wealth of existing GHQ-12 data.

Secondly, whilst it is clear that fit indices favour the proposed structure, it is important to establish that the proposed structure here offers more substantively in terms of predictive capacity than simpler structures. Recommendations for fit indices come from studies using

sample sizes far smaller than those used here (Bentler, 2007; Hu and Bentler, 1999). As such, in isolation there are multiple structures tested here which would be accepted as “adequate” under such criteria. Whilst the Bayesian fit statistics presented are more robust to sample size, more work is needed on the appropriateness of fit-statistic thresholds in the face of increasingly large sample sizes (Muthén and Asparouhov, 2010). Proposed structures must be evaluated not just numerically, but on a theoretical basis, identifying whether the more complex structure offers greater understanding. The mean absolute factor correlations give some indication of this in terms of substantive dissimilarity between constructs. However comprehensive evaluation of this requires the investigation of the predictors of the different constructs, demographically, geographically and socially to evidence whether they truly add to our understanding of different processes. As such, further work is needed using the decomposed constructs as responses with data for which the structure is validated.

Similarly, given the use of the GHQ-12 as an external validation instrument (e.g. Mukuria *et al.*, 2014), it is important to incorporate these more nuanced understandings in subsequent studies evaluating what is being captured by other existing or new metrics (e.g. Tennant *et al.*, 2007). Adopting this approach will allow researchers to go further than modelling similarities in manifest responses to understanding similarities between underpinning mental health processes across populations. One such avenue is to investigate if any of the structures identified here are associated with similar underpinning processes of more recently developed well-being metrics such as the Short Warwick-Edinburgh Mental Well-Being Scale (Stewart-Brown *et al.*, 2009). This will allow an empirical contribution to the debate on the contested relationship of mental illness and well-being (Westerhof and Keyes, 2010).

There are further limitations in the interpretation of the resulting constructs. It has been suggested in GHQ-12 literature that multidimensional factor structures are simply a product

of over-interpretation of spurious variance in negatively worded items (Hankins, 2008a). It is important to note the possibility of this being the case in this data, the unidimensional correlated-error model performs very well, given its brevity. It initially seems reasonable to infer from this the potential for multidimensionality being solely the result of phrasing, but only if one assumes unidirectional causality, i.e. items were grouped into positive and negative items at random, rather than based upon conceptually different measurements (Gnambs and Staufenbiel, 2018). Empirically these two scenarios would present identically, although it seems reasonable to assume greater likelihood of the latter.

Replicating this analysis across different temporal and spatial contexts is a clear avenue of further research. Whilst understanding the stability of the structure across contexts is undoubtedly important, it is also important to demonstrate its worth in further developing understanding of what is being captured by these measures in large scale surveys. It is important, therefore, to evaluate the insight gained by decomposing complex metrics by taking them as responses in spatial, social, structural and genetic epidemiological studies. It is clearly of greater benefit to social scientists in any analysis positing putative causal mechanisms to robustly link predictors with *underpinning processes* of mental health than with aggregate, unidimensional questionnaire responses.

In conclusion this study, despite the above limitations, is one of the first to combine Bayesian and ESEM methodology with large-scale survey data in the UK and has demonstrated the inferential benefits of doing so for research into population-level mental health determinants. Further research is needed to validate these findings, using data from wider contexts, and contextualise them against differing mental health responses.

References

- 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, 517–25. <https://doi.org/10.1177/1073191110388146>
- Andrich, D., Schoubroeck, L. Van, 1989. The General Health Questionnaire: a psychometric analysis using latent trait theory. *Psychol. Med.* 19, 469–485. <https://doi.org/10.1017/S0033291700012502>
- Asparouhov, T., Muthén, B., 2010. Bayesian analysis using Mplus: Technical implementation. Los Angeles Muthén Muthén 1–38.
- Asparouhov, T., Muthén, B., 2009a. Exploratory Structural Equation Modeling, *Structural Equation Modeling: A Multidisciplinary Journal*. <https://doi.org/10.1080/10705510903008204>
- Asparouhov, T., Muthén, B., Morin, A.J.S., 2015. Bayesian Structural Equation Modeling With Cross-Loadings and Residual Covariances: Comments on Stromeier et al. . *J. Manag.* 41, 1561–1577. <https://doi.org/10.1177/0149206315591075>
- Asparouhov, T., Muthén, B.O., 2009b. Exploratory Structural Equation Modeling, *Structural Equation Modeling: A Multidisciplinary Journal*. <https://doi.org/10.1080/10705510903008204>
- Baksheev, G.N., Robinson, J., Cosgrave, E.M., Baker, K., Yung, A.R., 2011. Validity of the 12-item General Health Questionnaire (GHQ-12) in detecting depressive and anxiety disorders among high school students. *Psychiatry Res.* 187, 291–6. <https://doi.org/10.1016/j.psychres.2010.10.010>
- Bentler, P.M., 2007. On tests and indices for evaluating structural models. *Pers. Individ. Dif.* 42, 825–829. <https://doi.org/10.1016/j.paid.2006.09.024>
- Brodersen, J., McKenna, S.P., Doward, L.C., et al., 2007. Measuring the psychosocial consequences of screening. *Health Qual. Life Outcomes* 5, 3. <https://doi.org/10.1186/1477-7525-5-3>
- Browne, M.W., 2001. An overview of analytic rotation in exploratory factor analysis. *Multivariate Behav. Res.* 36, 111–150. https://doi.org/10.1207/S15327906MBR3601_05
- Cai, N., Kendler, K.S., Flint, J., 2018. Minimal phenotyping yields GWAS hits of low specificity for major depression. *BioRxiv* 44, 300.
- Campbell, A., Knowles, S., 2007. A confirmatory factor analysis of the GHQ12 using a large Australian sample. *Eur. J. Psychol. Assess.* 23, 2–8. <https://doi.org/10.1027/1015-5759.23.1.2>
- Campbell, A., Walker, J., Farrell, G., 2003. Confirmatory factor analysis of the GHQ-12: can I see that again? *Aust. N. Z. J. Psychiatry* 37, 475–483. <https://doi.org/10.1046/j.1440-1614.2003.01208.x>
- Chen, F.F., 2007. Sensitivity of goodness of fit indexes to lack of measurement invariance. *Struct. Equ. Model. A Multidiscip. J.* 14, 464–504. <https://doi.org/10.1080/10705510701301834>
- Cheung, Y.B., 2002. A confirmatory factor analysis of the 12-item General Health Questionnaire among older people 739–744.
- Conway, J.M., Huffcutt, A.I., 2003. A Review and Evaluation of Exploratory Factor Analysis Practices in Organizational Research. *Organ. Res. Methods* 6, 147–168. <https://doi.org/10.1177/1094428103251541>

- 538 Creed, P.A., Evans, B.M., 2002. Personality, well-being and deprivation theory. *Pers. Individ. Dif.* 33,
539 1045–1054. [https://doi.org/10.1016/S0191-8869\(01\)00210-0](https://doi.org/10.1016/S0191-8869(01)00210-0)
- 540 Cronbach, L.J., 1951. Coefficient alpha and the internal structure of tests. *Psychometrika* 16, 297–334.
541 <https://doi.org/10.1007/BF02310555>
- 542 Davies, S., 2013. Annual Report of the Chief Medical Officer 2013, Public Health Mental Priorities:
543 Investing in the evidence 320.
- 544 Davies, S.C., 2018. Annual Report of the Chief Medical Officer, 2018: Health 2040 - Better Health Within
545 Reach.
- 546 Dunn, T.J., Baguley, T., Brunsden, V., 2014. From alpha to omega: A practical solution to the pervasive
547 problem of internal consistency estimation. *Br. J. Psychol.* 105, 399–412.
548 <https://doi.org/10.1111/bjop.12046>
- 549 Fabrigar, L.R., Wegener, D.T., MacCallum, R.C., Strahan, E.J., 1999. Evaluating the use of exploratory
550 factor analysis in psychological research. *Psychol. Methods* 4, 272–299.
551 <https://doi.org/10.1037//1082-989X.4.3.272>
- 552 Faris, R.E.L., Dunham, H.W., 1939. Mental disorders in urban areas: an ecological study of schizophrenia
553 and other psychoses. Univ. Chicago Press, Oxford, England.
- 554 Fernandes, H.M., Vasconcelos-Raposo, J., 2013. Factorial validity and invariance of the GHQ-12 among
555 clinical and nonclinical samples. *Assessment* 20, 219–29.
556 <https://doi.org/10.1177/1073191112465768>
- 557 French, D.J., Tait, R.J., 2004. Measurement invariance in the General Health Questionnaire-12 in young
558 Australian adolescents. *Eur. Child Adolesc. Psychiatry* 13, 1–7. [https://doi.org/10.1007/s00787-004-](https://doi.org/10.1007/s00787-004-0345-7)
559 0345-7
- 560 Gao, F., Luo, N., Thumboo, J., Fones, C., Li, S.-C., Cheung, Y.-B., 2004. Does the 12-item General Health
561 Questionnaire contain multiple factors and do we need them? *Health Qual. Life Outcomes* 2, 63.
562 <https://doi.org/10.1186/1477-7525-2-63>
- 563 Garrard, J., Rolnick, S.J., Nitz, N.M., Luepke, L., Jackson, J., Fischer, L.R., Leibson, C., Bland, P.C., Heinrich,
564 R., Waller, L.A., 1998. Clinical detection of depression among community-based elderly people with
565 self-reported symptoms of depression. *J. Gerontol. A Biol. Sci. Med. Sci.* 53, M92–M101.
- 566 Gelman, a, Gelman, a, Meng, X.-L., Meng, X.-L., Stern, H., Stern, H., 1996. Posterior predictive
567 assessment of model fitness via realized discrepancies. Vol.6, No.4. *Stat. Sin.* 6, 733–807.
568 <https://doi.org/10.1.1.142.9951>
- 569 Gnamb, T., Staufenbiel, T., 2018. The structure of the General Health Questionnaire (GHQ-12): two
570 meta-analytic factor analyses. *Health Psychol. Rev.* 12, 179–194.
571 <https://doi.org/10.1080/17437199.2018.1426484>
- 572 Goldberg, D., Bridges, K., 1987. Screening for psychiatric illness in general practice: the general
573 practitioner versus the screening questionnaire. *J. R. Coll. Gen. Pract.* 37, 15–18.
- 574 Goldberg, D., Gater, R., Sartorius, N., Ustün, T.B., Piccinelli, M., Gureje, O., Rutter, M., 1997. The validity
575 of two versions of the GHQ in the WHO study of mental illness in general health care. *Psychol. Med.*
576 27, 191–197.
- 577 Goldberg, D.P., Hillier, V.F., 1979. A scaled version of the General Health Questionnaire. *Psychol. Med.* 9,

578 139–145. <https://doi.org/10.1017/S0033291700021644>

579 Goldberg, D.P., Oldehinkel, T., Ormel, J., 1998. Why GHQ threshold varies from one place to another.
580 Psychol. Med. 28, 915–21. <https://doi.org/10.1017/S0033291798006874>

581 Gouveia, V. V, Barbosa, G.A., Oliveira Andrade, E. De, Carneiro, M.B., 2010. Factorial validity and
582 reliability of the General Health Questionnaire (GHQ-12) in the Brazilian physician population. Cad.
583 Saude Publica 26, 1439–1445. <https://doi.org/10.1590/S0102-311X2010000700023>

584 Graetz, B., 1991. Multidimensional properties of the general health questionnaire. Soc. Psychiatry
585 Psychiatr. Epidemiol. 132–138.

586 Hankins, M., 2008a. The reliability of the twelve-item general health questionnaire (GHQ-12) under
587 realistic assumptions. BMC Public Health 8, 355. <https://doi.org/10.1186/1471-2458-8-355>

588 Hankins, M., 2008b. Clinical Practice and Epidemiology The factor structure of the twelve item General
589 Health Questionnaire (GHQ-12): the result of negative phrasing? 8, 1–8.
590 <https://doi.org/10.1186/1745-0179-4-Received>

591 Hu, L., Bentler, P.M., 1999. Cutoff criteria for fit indexes in covariance structure analysis: Conventional
592 criteria versus new alternatives. Struct. Equ. Model. A Multidiscip. J. 6, 1–55.
593 <https://doi.org/10.1080/10705519909540118>

594 Hu, Y., Stewart-Brown, S., Twigg, L., Weich, S., 2007. Can the 12-item General Health Questionnaire be
595 used to measure positive mental health? Psychol. Med. 37, 1005–13.
596 <https://doi.org/10.1017/S0033291707009993>

597 Jöreskog, K.G., 1969. A general approach to confirmatory maximum likelihood factor analysis.
598 Psychometrika 34, 183–202. <https://doi.org/10.1007/BF02289343>

599 Kalliath, T.J., O'Driscoll, M.P., Brough, P., 2004. A confirmatory factor analysis of the General Health
600 Questionnaire-12. Stress Heal. 20, 11–20. <https://doi.org/10.1002/smi.993>

601 Khan, A., Mad Shah, I., Khan, F., Suhail, S., 2013. Reliability and validity assessment of 12 items general
602 health questionnaire (GHQ: 12) among Pakistani university teachers. World Appl. Sci. J. 24, 603–
603 608. <https://doi.org/10.5829/idosi.wasj.2013.24.05.13212>

604 Kiliç, C., Rezaki, M., Rezaki, B., Kaplan, I., Özgen, G., Saıduyu, A., Ozırk, M.O., 1997. General Health
605 Questionnaire (GHQ12 and GHQ28): Psychometric properties and factor structure of the scales in a
606 Turkish primary care sample. Soc. Psychiatry Psychiatr. Epidemiol. 32, 327–331.
607 <https://doi.org/10.1007/BF00805437>

608 Kish, L., 1954. Differentiation in Metropolitan Areas. Am. Sociol. Rev. 19, 388–398.
609 <https://doi.org/10.2307/2087457>

610 Lecrubier, Y., 2007. Widespread underrecognition and undertreatment of anxiety and mood disorders:
611 Results from 3 European studies. J. Clin. Psychiatry 68, 36–41.

612 Lynn, P., Knies, G., 2016. Understanding Society: The UK Household Longitudinal Study Waves 1-5 Quality
613 Profile. Inst. Soc.

614 MacCallum, R.C., Roznowski, M., Necowitz, L.B., 1992. Model modifications in covariance structure
615 analysis: The problem of capitilization on chance. Quant. Methods Psychol.
616 <https://doi.org/10.1037/0033-2909.111.3.490>

- 617 Marsh, H.W., Bailey, M., 1991. Confirmatory Factor Analyses of Multitrait-Multimethod Data: A
618 Comparison of Alternative Models. *Appl. Psychol. Meas.* 15, 47–70.
619 <https://doi.org/10.1177/014662169101500106>
- 620 Marsh, H.W., Hau, K.-T., Wen, Z., 2004. In Search of Golden Rule: Comment on Hypothesis Testing
621 Approaches to Setting Cutoff Value for Fit Indexes and Danger in Overgeneralizing Hu and Bentler's
622 (1999) Finding. *Struct. Equ. Model.* 11, 320–341. <https://doi.org/10.1207/s15328007sem1103>
- 623 Marsh, H.W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A.J.S., Trautwein, U., Nagengast, B., 2010. A
624 new look at the big five factor structure through exploratory structural equation modeling. *Psychol.*
625 *Assess.* 22, 471–491. <https://doi.org/10.1037/a0019227>
- 626 Marsh, H.W., Morin, A.J.S., Parker, P.D., Kaur, G., 2014. Exploratory structural equation modeling: an
627 integration of the best features of exploratory and confirmatory factor analysis. *Annu. Rev. Clin.*
628 *Psychol.* 10, 85–110. <https://doi.org/10.1146/annurev-clinpsy-032813-153700>
- 629 Marsh, H.W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A.J.S., Trautwein, U., 2009.
630 Exploratory Structural Equation Modeling, Integrating CFA and EFA: Application to Students'
631 Evaluations of University Teaching, *Structural Equation Modeling: A Multidisciplinary Journal*.
632 <https://doi.org/10.1080/10705510903008220>
- 633 Martin, A.J., 1999. Assessing the Multidimensionality of the 12-Item General Health Questionnaire.
634 *Psychol. Rep.* 84, 927–935.
- 635 Martin, C.R., Newell, R.J., 2005. The factor structure of the 12-item General Health Questionnaire in
636 individuals with facial disfigurement. *J. Psychosom. Res.* 59, 193–9.
637 <https://doi.org/10.1016/j.jpsychores.2005.02.020>
- 638 McDonald, R.P., 2005. Semiconfirmatory Factor Analysis: The Example of Anxiety and Depression. *Struct.*
639 *Equ. Model. A Multidiscip. J.* 12, 163–172. https://doi.org/10.1207/s15328007sem1201_9
- 640 McFall, S.L., 2011. Understanding Society - The UK household longitudinal study: Wave 1, 2009-2010,
641 User Manual. Colchester.
- 642 Mindell, J., Biddulph, J.P., Hirani, V., Stamatakis, E., Craig, R., Nunn, S., Shelton, N., 2012. Cohort profile:
643 The health survey for england. *Int. J. Epidemiol.* 41, 1585–1593. <https://doi.org/10.1093/ije/dyr199>
- 644 Mukuria, C., Rowen, D., Peasgood, T., Brazier, J., 2014. An empirical comparison of well-being measures
645 used in UK 1–55.
- 646 Muthén, B., Asparouhov, T., 2010. Bayesian SEM: A more flexible representation of substantive theory.
- 647 Padrón, A., Galán, I., Durbán, M., Gandarillas, A., Rodríguez-Artalejo, F., 2012. Confirmatory factor
648 analysis of the General Health Questionnaire (GHQ-12) in Spanish adolescents. *Qual. Life Res.* 21,
649 1291–8. <https://doi.org/10.1007/s11136-011-0038-x>
- 650 Patalay, P., Fitzsimons, E., 2016. Correlates of Mental Illness and Wellbeing in Children: 55.
651 <https://doi.org/10.1016/j.jaac.2016.05.019>
- 652 Penninkilampi-Kerola, V., Miettunen, J., Ebeling, H., 2006. A comparative assessment of the factor
653 structures and psychometric properties of the GHQ-12 and the GHQ-20 based on data from a
654 Finnish population-based sample. *Scand. J. Psychol.* 47, 431–40. <https://doi.org/10.1111/j.1467-9450.2006.00551.x>
- 656 Pevalin, D.J., 2000. Multiple applications of the GHQ-12 in a general population sample: an investigation

657 of long-term retest effects. *Soc. Psychiatry Psychiatr. Epidemiol.* 35, 508–512.
658 <https://doi.org/10.1007/s001270050272>

659 Picardi, a, Abeni, D., Pasquini, P., 2001. Assessing psychological distress in patients with skin diseases:
660 reliability, validity and factor structure of the GHQ-12. *J. Eur. Acad. Dermatology Venereol.* 15, 410–
661 417. <https://doi.org/10.1046/j.1468-3083.2001.00336.x>

662 Politi, P.L., Piccinelli, M., Wilkinson, G., 1994. Reliability, validity and factor structure of the 12-item
663 General Health Questionnaire among young males in Italy. *Acta Psychiatr. Scand.* 90, 432–437.
664 <https://doi.org/10.1111/j.1600-0447.1994.tb01620.x>

665 Propper, C., Jones, K., Bolster, A., Burgess, S., Johnston, R., Sarker, R., 2005. Local neighbourhood and
666 mental health: evidence from the UK. *Soc. Sci. Med.* 61, 2065–83.
667 <https://doi.org/10.1016/j.socscimed.2005.04.013>

668 Romppel, M., Braehler, E., Roth, M., Glaesmer, H., 2013. What is the General Health Questionnaire-12
669 assessing?: Dimensionality and psychometric properties of the General Health Questionnaire-12 in
670 a large scale German population sample. *Compr. Psychiatry* 54, 406–413.
671 <https://doi.org/10.1016/j.comppsy.2012.10.010>

672 Sánchez-López, M.D.P., Dresch, V., 2008. The 12-Item General Health Questionnaire (GHQ-12): reliability,
673 external validity and factor structure in the Spanish population. *Psicothema* 20, 839–43.

674 Schmitt, T. a., 2011. Current Methodological Considerations in Exploratory and Confirmatory Factor
675 Analysis. *J. Psychoeduc. Assess.* 29, 304–321. <https://doi.org/10.1177/0734282911406653>

676 Schmitz, N., Kruse, J., Tress, W., 1999. Psychometric properties of the General Health Questionnaire
677 (GHQ-12) in a German primary care sample. *Acta Psychiatr. Scand.* 100, 462–468.
678 <https://doi.org/10.1111/j.1600-0447.1999.tb10898.x>

679 Shevlin, M., Adamson, G., 2005. Alternative factor models and factorial invariance of the GHQ-12: a large
680 sample analysis using confirmatory factor analysis. *Psychol. Assess.* 17, 231–236.
681 <https://doi.org/10.1037/1040-3590.17.2.231>

682 Smith, A.B., Fallowfield, L.J., Stark, D.P., Velikova, G., Jenkins, V., 2010. A Rasch and confirmatory factor
683 analysis of the general health questionnaire (GHQ)--12. *Health Qual. Life Outcomes* 8, 45.
684 <https://doi.org/10.1186/1477-7525-8-45>

685 Smith, A.B., Oluboyede, Y., West, R., Hewison, J., House, A.O., 2013. The factor structure of the GHQ-12:
686 the interaction between item phrasing, variance and levels of distress. *Qual. Life Res.* 22, 145–52.
687 <https://doi.org/10.1007/s11136-012-0133-7>

688 Stewart-Brown, S., Tennant, A., Tennant, R., Platt, S., Parkinson, J., Weich, S., 2009. Internal construct
689 validity of the Warwick-Edinburgh Mental Well-being Scale (WEMWBS): a Rasch analysis using data
690 from the Scottish Health Education Population Survey. *Health Qual. Life Outcomes* 7, 15.
691 <https://doi.org/10.1186/1477-7525-7-15>

692 Suzuki, H., Kaneita, Y., Osaki, Y., Minowa, M., Kanda, H., Suzuki, K., Wada, K., Hayashi, K., Tanihata, T.,
693 Ohida, T., 2011. Clarification of the factor structure of the 12-item General Health Questionnaire
694 among Japanese adolescents and associated sleep status. *Psychiatry Res.* 188, 138–46.
695 <https://doi.org/10.1016/j.psychres.2010.10.025>

696 Sweeting, H., Young, R., West, P., 2009. GHQ increases among Scottish 15 year olds 1987-2006. *Soc.*
697 *Psychiatry Psychiatr. Epidemiol.* 44, 579–86. <https://doi.org/10.1007/s00127-008-0462-6>

- Tait, R.J., French, D.J., Hulse, G.K., 2003. Validity and psychometric properties of the General Health Questionnaire-12 in young Australian adolescents. *Aust. N. Z. J. Psychiatry* 37, 374–381. <https://doi.org/10.1046/j.1440-1614.2003.01133.x>
- Tennant, R., Hiller, L., Fishwick, R., Platt, S., Joseph, S., Weich, S., Parkinson, J., Secker, J., Stewart-Brown, S., 2007. The Warwick-Edinburgh Mental Well-being Scale (WEMWBS): development and UK validation. *Health Qual. Life Outcomes* 5, 63. <https://doi.org/10.1186/1477-7525-5-63>
- Thomson, R.M., Katikireddi, S.V., 2018. Mental health and the jilted generation: Using age-period-cohort analysis to assess differential trends in young people’s mental health following the Great Recession and austerity in England. *Soc. Sci. Med.* 214, 133–143. <https://doi.org/10.1016/j.socscimed.2018.08.034>
- Vanheule, S., Bogaerts, S., 2005. The factorial structure of the GHQ-12. *Stress Heal.* 21, 217–222. <https://doi.org/10.1002/smi.1058>
- Verheij, R., 1996. Explaining urban-rural variations in health: a review of interactions between individual and environment. *Soc. Sci. Med.* 42, 923–935.
- Wang, L., Lin, W., 2011. Wording effects and the dimensionality of the General Health Questionnaire (GHQ-12). *Pers. Individ. Dif.* 50, 1056–1061. <https://doi.org/10.1016/j.paid.2011.01.024>
- Weich, S., Holt, G., Twigg, L., Jones, K., Lewis, G., 2003. Geographic Variation in the Prevalence of Common Mental Disorders in Britain: A Multilevel Investigation. *Am. J. Epidemiol.* 157, 730–737. <https://doi.org/10.1093/aje/kwg035>
- Werneke, U., Goldberg, D.P., Yalcin, I., Üstün, B.T., 2000. The stability of the factor structure of the General Health Questionnaire. *Psychol. Med.* 30, 823–829. <https://doi.org/10.1017/S0033291799002287>
- Westerhof, G.J., Keyes, A.C.L.M., 2010. Mental Illness and Mental Health: The Two Continua Model Across the Lifespan 110–119. <https://doi.org/10.1007/s10804-009-9082-y>
- Winefield, H.R., Goldney, R.D., Winefield, A.H., Tiggemann, M., 1989. The General Health Questionnaire: Reliability and validity for Australian youth. *Aust. N. Z. J. Psychiatry* 23, 53–58. <https://doi.org/10.3109/00048678909062592>
- World Health Organization, 2013. Mental Health Action Plan 2013-2020. *WHO Libr. Cat. DataLibrary Cat.* Data 1–44. [https://doi.org/ISBN 978 92 4 150602 1](https://doi.org/ISBN%20978%2092%204%20150602%201)
- Worsley, A., Gribbin, C.C., 1977. A factor analytic study on the twelve item general health questionnaire. *Aust. N. Z. J. Psychiatry* 11, 260–72. <https://doi.org/10.3109/00048677709159577>
- Ye, S., 2009. Factor structure of the General Health Questionnaire (GHQ-12): The role of wording effects. *Pers. Individ. Dif.* 46, 197–201. <https://doi.org/10.1016/j.paid.2008.09.027>