University of Dundee

University of Dundee

Calibrating well-being, quality of life and common mental disorder items

Boehnke, Jan; Croudace, Tim J.

[Link to publication in Discovery Research Portal](#)

Calibrating well-being, quality of life and common mental disorder items: psychometric epidemiology in public mental health research

Jan R. Böhnke and Tim J. Croudace

Background
The assessment of 'general health and well-being' in public mental health research stimulates debates around relative merits of questionnaire instruments and their items. Little evidence regarding alignment or differential advantages of instruments or items has appeared to date.

Aims
Population-based psychometric study of items employed in public mental health narratives.

Method
Multidimensional item response theory was applied to General Health Questionnaire (GHQ-12), Warwick-Edinburgh Mental Well-being Scale (WEMWBS) and EQ-5D items (Health Survey for England, 2010–2012; $n = 19\,290$).

Results
A bifactor model provided the best account of the data and showed that the GHQ-12 and WEMWBS items assess mainly the same construct. Only one item of the EQ-5D showed

relevant overlap with this dimension (anxiety/depression). Findings were corroborated by comparisons with alternative models and cross-validation analyses.

Conclusions
The consequences of this lack of differentiation (GHQ-12 *v.* WEMWBS) for mental health and well-being narratives deserves discussion to enrich debates on priorities in public mental health and its assessment.

Declaration of interest
T.J.C. reports: grants from GL Assessment (2008–2011) and a personal fee from GL Assessment outside the submitted work. GL Assessment market and distribute the General Health Questionnaire.

Psychological assessment of populations in the form of self-report questionnaires and patient-reported outcome measures play a key role in psychiatric research, clinical epidemiology and in public (mental) health. The ability of questionnaire instruments to reliably assess multiple aspects of health has been questioned by recent research,[1–4] and the finding that one dominant dimension explains a large proportion of the variability in many instruments has consequences for debates about the (psychometric) epidemiology of population health and well-being.[5,6] How do we accurately assess the psychological well-being of a nation, be it once or repeatedly? With which sets of instruments or items? Many governments are exploring strategies to assess population well-being using both established or experimental measures to inform or evaluate policy with some broader notions than economic measures of social progress.[7,8] We report the first epidemiological study with contemporary psychometric techniques of the (dis)similarities of instruments that appear in many archived UK Government surveys and that are also discussed as future measures: Goldberg's General Health Questionnaire (GHQ) items[9] and the more recently developed Warwick-Edinburgh Mental Well-being Scale (WEMWBS).[10] This article reports on a psychometric evaluation of whether (or to which degree) the two instruments assess different constructs in representative UK population samples. The analysis also includes the items of the EQ-5D[11] to corroborate that the chosen analytic methods can detect differences between items that address related, but sufficiently different aspects of subjective health.

Method

Participants

We studied respondents from three waves of the Health Survey for England (HSE, 2010,[12] 2011,[13] 2012[14]). Our analyses included respondents older than 16 who were scheduled for the survey's self-completion module with the three relevant instruments. Since the HSE is run on a household basis, we selected up to two respondents from the same household (one for the estimation sample and a second, different one for the validation sample[15]). Table 1 provides information on the number of such respondents available for analysis.

For efficiency, the survey responses were pooled across all three data-sets (total $n = 19\,290$). The use of partly incomplete data (since the 12-item GHQ (GHQ-12) was not collected in 2011) is an instance of a 'planned missing data' design for which standard solutions from the analytical literature on integrative data analysis exist.[16–18] Random sample splits generated one estimation sample ($n = 9669$) – in which the statistical models were optimised – and a validation sample ($n = 9621$) that we evaluated to test whether our best-fitting model was not purely a result of overfitting to one data-set, a possibility we hoped to avoid.[19]

Measures

The GHQ[9] is a self-report questionnaire aimed at screening for diagnosable levels of psychological distress in healthcare settings.

Table 1 Number of respondents in each wave of the Health Survey for England (HSE) and number of respondents that had at least one response on the respective instrument

	Total respondents			Respondents with at least one response on:		
	n	Women, %	Age, years: mean (s.d.)	EQ-5D, *n*	12-item General Health Questionnaire	Warwick-Edinburgh Mental Well-being Scale
HSE 2010	7255	56.9	50.52 (18.51)	7234	7223	7153
HSE 2011	7246	56.5	49.91 (18.25)	7182	?	7163
HSE 2012	4789	56.7	52.32 (17.91)	4742	4739	4779

The GHQ-12 has been shown to provide excellent reliability and validity as a screen for risk of common mental disorder. It is popular because of its brevity and its frequent use provides comparability across studies.[20] An important feature of the GHQ-12 is its inclusion of six positively and six negatively phrased items that have different verbal descriptors of their response options. Items are laid-out so that higher and right-most responses indicate greater distress. Whether these items measure one distress dimension and a method factor (correcting for wording effects) or two correlated dimensions is a topic of debate[15,21–23] and will be addressed by our statistical analysis (see below and online supplement DS1).

The WEMWBS aims at assessing mental well-being and its emotional, cognitive and psychological aspects.[10] The WEMWBS was designed to provide a short (14 items) and reliable assessment with solely positively phrased items. All items have five response options, anchored from 'None of the time' to 'All of the time'. Higher scores indicate 'well-being'. Studies have reported it to be a reliable instrument,[24] largely unidimensional,[25] with possible value in predicting health-related behaviours.[26]

We included the EQ-5D (EuroQol)[11,27] as a credible set of items to gauge the sensitivity of our analytic strategy to the presence of discriminant validity. The five items each assess a dimension of health-related quality of life (mobility, self-care, usual activities, pain/discomfort and anxiety/depression). Responses to these items are captured with three response options and item-specific anchors (for example, mobility: 'no problems walking about', 'some problems walking about', 'confined to bed'). Although researchers increasingly use the score of the EQ-5D as a dimensional indicator of health-related quality of life, it is important to note that the items were not developed as indicators of a latent dimension and are used in this study solely to explore discriminant validity. Instead, the EQ-5D was developed and validated to order 243 different health states, i.e. response patterns, according to their relative valuations in population samples.[11]

Statistical analysis

Item response factor models can be used (a) to establish the number of latent variables needed to explain the responses to the $k = 31$ items of the three instruments and (b) to investigate relationships between latent variables.[28] All models were estimated with Mplus 7.11[29] using full information maximum likelihood (FIML) estimation.[30] To explore the dimensional structure of the questionnaires[4,31] we assessed first the fit of a one-dimensional model (model 1) to test whether the instruments all measure the same latent variable (for example, well-being). We then assessed an exploratory correlated factor model with three factors (model 2) to test whether the instruments are separable and therefore assess different correlated constructs. The last set of models (model 3 and model 4) were bifactor models.[32] Bifactor models assume that all items load on one general dimension, but that there are remaining sources of covariation because of common

characteristics of certain item sets. Model fit was assessed based on information criteria that identify the best description of the data while penalising for model complexity.[33] The best-fitting model was cross-validated in the validation sample.

To evaluate the relative importance of the identified dimensions, we first calculated (partial) test information functions, i.e. the inverse of the measurement error. These functions provide a graphical evaluation of how accurately different sets of items assess the latent factor (relative efficiency).[34,35] The second strategy aimed at assessing the amount of score variance as a result of each factor,[32] i.e. assessing the reliability of the general factor and the three instrument-specific factors. To this end, we evaluated omega (ω) coefficients. Of these, ω (without subscript) assesses the share of score variance as a result of all factors of the bifactor model taken together (general, specific, method); and omega-h (ω_h) in turn only assesses the share of score variance as a result of the general factor, i.e. the reliability of the general factor alone. We also assessed ω coefficients for the three instruments (GHQ-12, WEMWBS, EQ-5D) individually, which provide the reliability of each instrument as derived from the bifactor model, as well as omgea-s (ω_s) coefficients, which assess the specific measurement quality of each instrument, when the general factor is partialled out.[36]

Results

For comprehensiveness, factor analyses for individual instruments appear in online supplement DS1. These reveal that one dimension explained most of the covariance in responses to both the GHQ-12 and the WEMWBS' 14 items. The model fit for each instrument was further optimised by adding a so-called 'method factor' for the negatively phrased items of the GHQ-12 and a factor for items dealing with social contacts and interests in the WEMWBS. These factors were retained in the following analyses. A confirmatory bifactor model with specific factors for the three instruments as well as two factors for wording effects (GHQ-12 items) and social/interest items (WEMWBS; model 4) was the best description of the data while penalising for complexity (Table 2) as indicated by both Bayesian information criterion (BIC) and adjusted BIC (BIC$_{adj}$, lowest values of all estimated models[33]). We discuss the results of three models in more detail: unidimensional, exploratory and bifactor.

Unidimensional

The unidimensional model (model 1; Table 2) assumes that one latent factor causes the responses to all three instruments. The items of both GHQ-12 and WEMWBS show high loadings on this factor (the lowest loading was $r = 0.54$; online Table DS1). Based on a minimal loading[37] of $r = 0.40$ three of the EQ-5D items load relevantly on this factor: anxiety/depression ($r = 0.71$), self-care ($r = 0.54$) and usual activities ($r = 0.47$). Compared with the mean loading of 0.68 (s.d. = 0.10) of the GHQ-12 and WEMWBS items

Table 2 Information criteria for the factor models with all items across the three instruments (12-item General Health Questionnaire, Warwick-Edinburgh Mental Well-being Scale and EQ-5D) in the estimation sample[a]

	Model 1: confirmatory one factor	Model 2: exploratory three factor, geomin-rotated	Model 3: confirmatory bifactor[b]	Model 4: confirmatory bifactor with two method factors[c]
Log likelihood (LL)	−209 811	−19 8980	−199 699	−198 689
Number of parameters (n(P))	133	192	164	175
Bayesian information criterion (BIC)	420 843	399 722	400 903	398 984
Adjusted BIC	420 420	399 111	400 381	398 428

a. $n = 9661$; estimator: full information maximum likelihood.
b. This model estimates only the general factor and three specific factors for the three instruments and was only estimated to test whether the addition of the method factors was necessary in model 4; since model 4 showed better fit, this model is not further discussed.
c. See Fig. 1 for details.

on this general factor, the latter two items load only marginally on this factor.

Exploratory

The exploratory model (model 2; Table 2) investigates whether three correlated latent constructs are present. The resulting factors are moderately correlated (see footnote to online Table DS1) and the first factor is highly loaded by items from the EQ-5D but only moderately so by some items from the other instruments. The second and third factors mirror the structure of the GHQ-12 and WEMWBS, with the anxiety/depression item of the EQ-5D loading with the GHQ-12 items (on the second factor).

The results from both the unidimensional and exploratory models indicate that instruments developed for very different purposes (GHQ-12 for screening; WEMWBS for mental well-being) share a considerable amount of variance and overlap in their measured range of distress or well-being.

Bifactor

The bifactor model (Fig. 1; model 4, Table 2) includes a (general) factor for all items, three (specific) factors for each instrument and also the two method factors that were identified in the single scale analyses (see online supplement DS1). The GHQ-12 items show equal or higher loadings on the general factor than expected from the results of the earlier unidimensional model (online Table DS1). The fact that no item shows higher loadings on the specific factors for negative wording or the scale itself indicates that the general factor is a reliable description of the response process for these items: the GHQ-12 items in effect define the metric of this (general/dominant) factor and variations along this continuum are anchored by their item content.

The items of the WEMWBS show lower loadings on this general factor than expected from the unidimensional solution

and its left-over covariation is captured both by the specific factor for the instrument and the factor for the social/interest items. Both of these specific factors show higher loadings for most items than on the general factor, i.e. these items tap into aspects beyond the general factor (i.e. detectable multidimensionality beyond the established GHQ-12 metric).

The anxiety/depression item of the EQ-5D aligns with the GHQ-12 metric, again 'standing off' from the other four items. After extraction of the general factor, the self-care and usual-activities items align with the other two non-mental-health items.

Cross-validation

To examine the robustness of our results, we treated the parameters (loadings and category thresholds) from the estimation sample as fixed and assessed how well they fitted the data of our validation sample.[4,19] Model 4 described the sample better than the closest runner-up, model 3 without the method factors (Table 3). Both models were re-estimated with weighted least squares means and variance adjusted (WLSMV) in the estimation sample.[38] The results were structurally very similar to the FIML solution and they were also checked in the validation sample. This procedure allowed for the estimation of typical fit indices for structural equation models that indicated good fit (Table 3) and also a model comparison test between the two solutions could be estimated that was highly significant ($\chi^2 = 1123.03$, d.f. $= 2$, $P < 0.001$). All indications are therefore supportive of the modelling claims and support the results found in the original analysis.

Information across the measurement range

All items (except four items of the EQ-5D) assess one strong dimension, making it of interest to understand how well the different items span the psychometric measurement range of this

Table 3 Model fit for cross-validation in the validation sample

	All parameters fixed		Free variances	
	Model 3:[b] bifactor	Model 4: full bifactor	Model 3:[b] bifactor	Model 4: full bifactor
Log likelihood (LL)	−198 233	−197 031	−198 233	−197 010
Number of parameters (n(P))	0	0	4	6
Bayesian information criterion (BIC)	396 468	394 061	396 502	394 074
Adjusted BIC	396 468	394 061	396 490	394 055
Root mean square error of approximation (90% CI)[c]	–	–	0.035 (0.034–0.035)	0.032 (0.031–0.033)
Tucker–Lewis index/comparative fit index[c]	–	–	0.98/0.98	0.98/0.99

a. $n = 9614$–9621 not completely missing responses.
b. This model estimates only the general factor and three specific factors for the three instruments and was only estimated to test whether the addition of the method factors was necessary in model 4.
c. Available only for model estimates from weighted least squares means and variance adjusted estimation (WLSMV) with n(P) > 1.
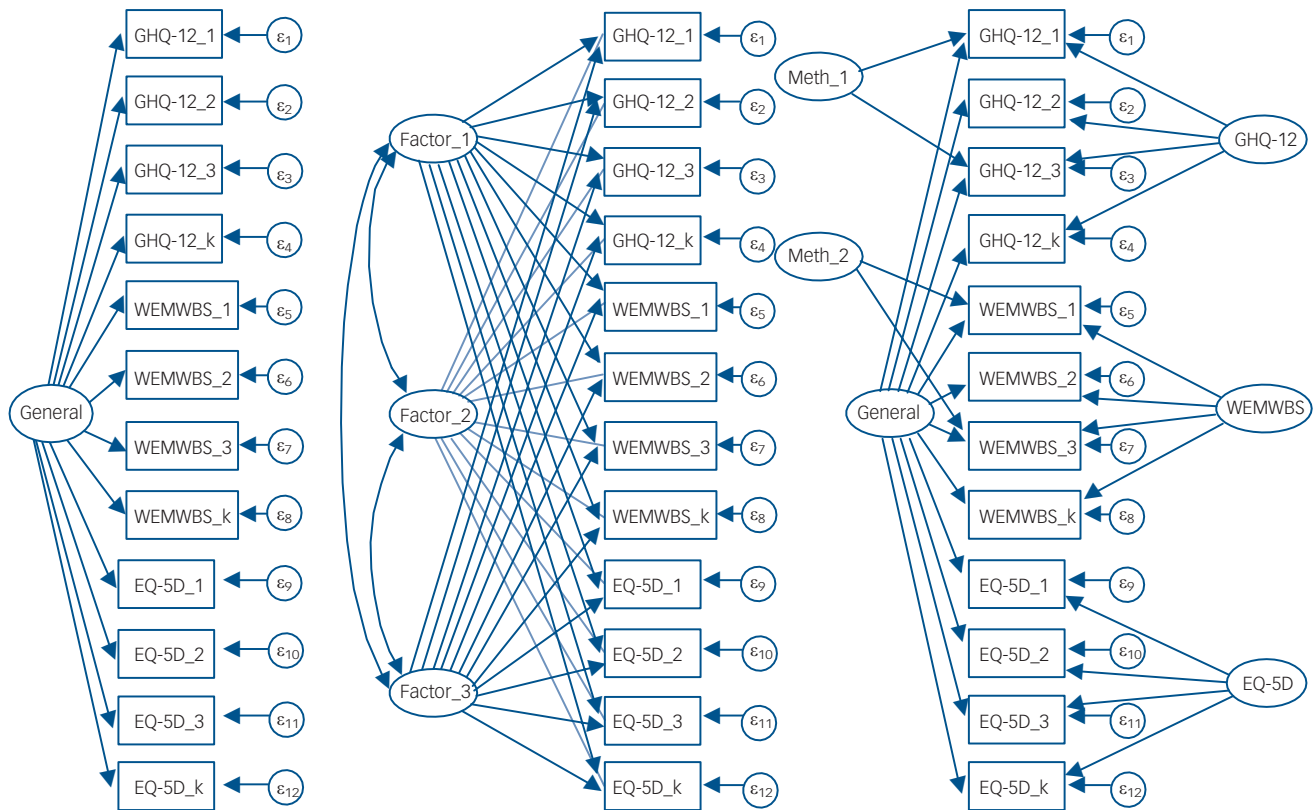
Fig. 1 Schematic representation of models 1, 2, and 4 (from left to right).

Arrows indicate loadings of an observed item (box) from the three instruments (12-item General Health Questionnaire (GHQ-12), Warwick-Edinburgh Mental Well-being Scale (WEMWBS), EQ-5D) on a latent variable (circle); Meth_1 indicates the method factor for wording of the GHQ-12 items; Meth_2 indicates the method factor for interest/social items of the WEMWBS items; model 3 is equivalent to model 4 without either of these method factors.

new factor. Item response theory-type bifactor analyses allow us to assess the range of the latent factor on which the item sets provide information as well as their relative contribution (relative efficiency[34,35]). The so-called information function (Fig. 2)[39] shows the inverse of the measurement error at each estimated latent score value (x-axis), hence more information indicates higher measurement precision for a given score level. Multiple lines offer direct comparisons of scales, with the pale blue line illustrating the information provided by the $k = 31$ items of the three instruments together. This information function has its peak between the sample mean (at '0' in Fig. 2) and a value almost two standard deviations below it. This is the region on which all items together provide most of the information and the standard error of measurement is the smallest. Given that this is a general population sample, the items on the whole span levels of (mild) distress.

The dark blue lines present how each instrument contributes separately to the measurement of the general factor ('partial information function'). The GHQ-12 has a higher information function (solid dark blue line) than the WEMWBS (dark blue dashed line) up to one standard deviation above the population mean. Beyond that, both instruments provide about the same amount of information (have similar accuracy). Since only one item of the EQ-5D (dark blue dotted line) loads relevantly on the general factor, it contributes only little to the measurement accuracy compared with the other 26 items.

Reliabilities

The $\omega = 0.98$ (online Table DS1) indicated that nearly all of the variance observed across the scores of the $k = 31$ items could be

attributed to the six sources that we defined in the bifactor model (i.e. the latent factors depicted in Fig. 1). The $\omega_h = 0.81$ for the general factor alone indicates still a reasonable reliability for the joint dimension across all items, but the difference between these two numbers shows that a substantial proportion of the score variance was as a result of the five specific factors of the model. Although the $\omega = 0.96$ across all GHQ-12 items indicates a high reliability, extracting the variance in GHQ-12 scores as a result of the general factor reduces this to an $\omega_s = 0.11$ – the GHQ-12 items contain no specific variance over and above the general factor. For the WEMWBS ($\omega = 0.95$; $\omega_s = 0.44$) and especially the EQ-5D ($\omega = 0.94$; $\omega_s = 0.61$) the results are less dramatic: both assess at least to some degree a specific dimension.

Discussion

To our knowledge this is the first study to comprehensively test the dimensionality of three instruments used to inform public mental health policy in the general population that also quantifies how differently they function in such a sample using modern psychometric methods.[6] Three general population samples and an established approach to merge multiple data sources were applied and the results showed that the GHQ-12 and the WEMWBS share one dimension and their items rather assess health states below the general population mean than above. The latter would be required to indicate a focus on positive states, and to potentially extend the continuum into 'well-being'. The finding of a strong general factor converges with other findings[3,4,31,39–42] and underlines the necessity to test whether alternatively developed and variously titled measures actually provide independent information.[1,6,43]
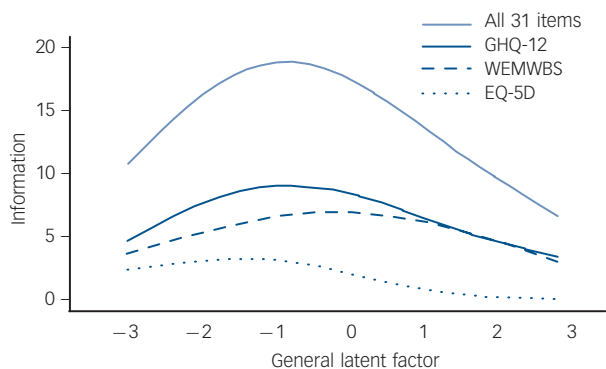
Fig. 2 Test information function for all 31 items of the three instruments (pale blue solid line) and partial test information functions for the items of the 12-item General Health Questionnaire (GHQ-12, dark blue solid line), the Warwick-Edinburgh Mental Well-being Scale (WEMWBS, dark blue dashed line) and the EQ-5D (dark blue dotted line).

The latent trait is the general dimension from the full bifactor model. The latent trait values are standardised with s.d. = 1 and '0' indicates the overall population mean on the general factor.

It is not possible to conclude this from instrument/item labelling or correlations between instrument scores alone, although that is often argued.

The bifactor modelling approach that we applied is a particularly powerful test for multidimensionality. For two instruments to assess different constructs the loadings on the general factor should be low at least for one instrument and the corresponding specific factor should show relevant loadings for all items of that scale. Both the GHQ-12 and the WEMWBS fail this test and for both instruments more than 50% of the common variance is because of the general dimension.[32] Although this finding questions the differential validity of the two instruments, on a more positive note it shows that the item responses from both instruments can be translated into each other since they are strongly connected by the general dimension.[18,44] Our results show that this is not advisable for the EQ-5D: it shares only a comparatively small amount of variation with the items of the other two instruments. This result also shows that the applied approach can identify differences between items that might be used in a similar setting, but that are geared at assessing different aspects of health and well-being. Therefore, the overlap between the GHQ-12, the WEMWBS and the single item of the EQ-5D is unlikely to be (only) a consequence of the statistical procedure.

Implications for well-being research and public mental health

It has been argued that a deficit-oriented perspective on the distress derived from symptoms of mental disorders should be complemented by a positive perspective emphasising 'well-being',[7,45,46] which has been a central feature in the narratives offered by proponents of positive psychology.[47] It appears closer to the broad definition of 'health' offered and endorsed by the World Health Organization.[48] Scholars working from the evidence base on mental disorders and psychiatric distress have argued that in populations shifting (mental) well-being to be more positive would decrease the prevalence and incidence of psychopathological symptoms or even disorder.[6,49] And last but not least, many governments consider well-being of their populations as a core assessment beyond gross domestic product, as a wider measure of economic and social progress.[7,8] One recent (but perhaps not definitive) consultation on this topic in the UK ended with the

WEMWBS being favoured over the GHQ-12: the 'GHQ-12 is a well validated and well used measure of mental health, however it is a screening instrument of mental illness, not a measure of mental wellbeing'.[50]

Whether two instruments developed for different purposes actually measure different constructs, is an empirical question.[1,43] Our research highlights that the latent constructs assessed by two instruments overlap to a far greater extent than expected from the consultation statement. It was well known that responses to these two instruments are correlated,[7,24,51] but our study shows compellingly that in the UK population all items of the two instruments share a common dimension, which explains a considerable share of their score variance.

The most important test for superiority of one instrument above the other is still missing: to establish whether one of the instruments is actually a better predictor of an agreed gold-standard of what 'well-being' really is. This remains a difficult challenge, but one not unique to well-being research in public mental health. It also applies to many social science concepts and often to 'caseness' in psychiatry. In our proposed framework the general factor being more highly correlated with the criterion would be conclusive evidence that the common part across items is closest to the criterion (for example, 'well-being'). If instead one of the specific latent scores is actually a better predictor than the other factor scores and/or the general factor, this would be evidence that one instrument is a more useful representation of the construct in question than the other.

We feel that there is still a large gap in the existing knowledge base and perhaps lack of consensus about (a) the relative independence of the variety of instruments to measure well-being, (b) the relative predictive power for any target or gold-standard reference that should be assessed and (c) the nature of this gold-standard criterion. If self-report assessments shall take on the task of assessing well-being in the population to guide and evaluate the effects of policy, we currently see a pressing need for research regarding all these aspects.[6] Integrating our proposed framework into standard investigations of the predictive value of a collection of items (for example, health risks or behaviours[26,52]) might help to close this gap.

If well-being is in fact a multidimensional construct,[7,46,53] the finding of a strong general factor in a small set of items is not surprising. Best psychometric practice would construe a scale for every aspect of well-being.[54] Aggregating across a range of different (sub)constructs of well-being (for example, mastery, personal relationships, hedonic aspects[7]) by building a single scale (i.e. a single score) across heterogeneous indicators may lead to the situation that only an evaluation of 'negativity' is the underlying commonality: most respondents can agree when they are lacking well-being, but the positive end-points might then differ across individuals to the extent that a unidimensional assessment of these is difficult to obtain.[55]

Strengths and limitations

The use of three UK-representative survey data-sets allowed a robust assessment of the research question. The large combined size enabled us to split the sample and to test the structure on a statistically independent sample. The use of current psychometric methodology enabled a detailed assessment of the items, the instruments and their scores. And although the study contained some exploratory elements, it was largely a confirmatory test of the differences between instruments that we might expect to appear in the evidence base to guide implementation of well-being assessments.

The greatest shortcoming is the lack of an external reference criterion of well-being to test its differential association with the

GHQ-12 and the WEMWBS. A second limitation addresses the meaning of the general factor: although it is convenient to assume that it captures the common substantive variation across instruments, it instead could capture mainly methodological artefacts across the three instruments.[56] To clarify, we note that we were not able to test this in the current data-sets and so it remains another pressing research hypothesis for the use of self-report questionnaire data in general health surveys.[5] A third limitation is that we deliberately did not give new or preferred definitions of what well-being is. We used common instruments that have been discussed as indicators or measures of public mental health.[7,50] The lack of finding major difference between measures can be for several reasons: (a) the respondents not being able to differentiate, (b) the instruments actually assessing one factor, or (c) a lack of theoretical differentiation when constructing the questionnaires. Fourth, we used what we consider 'short' instruments – those already deployed in surveys because of their brevity. We note that these are considerably shorter than they were at prior stages in their development and evaluation phases[9,10,57] and analyses of the long versions might come to different conclusions with regard to the importance of each of our factors.

Finally, we focused only on the aspect of similarity/dissimilarity of these measures. We did not consider other important aspects relevant to the assessment of population well-being such as the fairness of these instruments across regions or demographic groups in the UK, which would be needed to provide assessments that can reflect the state of well-being in a vibrant and fair society.[5] We think that this would distract from the primary message of the research we report here, which is that a much more nuanced discussion about well-being as an epidemiological dimension has to be held. The next step would then be to ensure that the chosen item collection allows a fair assessment.

In conclusion, contemporary psychometric methods and data integration offer a promising direction for future research innovations. Inherent in this perspective is our discouragement to other researchers to think only about testing or deploying existing single instruments, instead shifting the focus to testing 'collections of items': (a) to identify items that enable an optimal comparison on one general dimension (the general factor in our application), (b) to cover as many domain aspects as necessary in public and academic opinion (specific factors in this case) as well as (c) selecting those items that are maximally predictive of an external gold-standard of well-being. Collections of items and scales should be used, since there are no good scales but only good items for specific purposes.[5,58] And for assessing and improving the public's mental health and well-being the best items available should be chosen. Applying the proposed framework, items would also become linkable with different instruments across samples and studies.[2,59] This would be a truly contemporary psychometric epidemiological perspective on the well-being of populations.

Jan R. Böhnke, Dr. rer. nat., Department of Health Sciences and Hull York Medical School, Mental Health and Addiction Research Group, University of York, York; **Tim J. Croudace**, PhD, MSc, BSc(Hons), DipAppPsych, School of Nursing & Health Sciences, University of Dundee, Dundee, UK

Correspondence: Jan R. Böhnke, Department of Health Sciences and Hull York Medical School, Mental Health and Addiction Research Group, University of York ARRC Area 4, Department of Health Sciences, York YO10 5DD, UK. Email: jan.boehnke@york.ac.uk

References

1 Reininghaus U, Priebe S. Measuring patient-reported outcomes in psychosis: conceptual and methodological review. *Br J Psychiatry* 2012; **201**: 262–7.

2 Wahl I, Löwe B, Bjorner JB, Fischer F, Langs G, Voderholzer U, et al. Standardization of depression measurement: a common metric was developed for 11 self-report depression measures. *J Clin Epidemiol* 2014; **67**: 73–86.

3 Cohen RM, Greenberg JM, IsHak WW. Incorporating multidimensional patient-reported outcomes of symptom severity, functioning, and quality of life in the individual burden of illness index for depression to measure treatment impact and recovery in MDD. *JAMA Psychiatry* 2013; **70**: 343–50.

4 Böhnke JR, Lutz W, Delgadillo J. Negative affectivity as a transdiagnostic factor in patients with common mental disorders. *J Affect Disord* 2014; **166**: 270–8.

5 Böhnke JR, Croudace TJ. Factors of psychological distress: clinical value, measurement substance, and methodological artefacts. *Soc Psychiatry Psychiatr Epidemiol* 2015; **50**: 515–24.

6 Mehta N, Croudace T, Davies DSC. Public mental health: evidenced-based priorities. *Lancet* 2015; **385**: 1472–5.

7 Stewart-Brown S. Defining and measuring mental health and wellbeing. In *Public Mental Health; Global Perspectives* (eds L Knifton, N Quinn): 33–42. Open University Press, 2013.

8 Hicks S, Tinkler L, Allin P. Measuring subjective well-being and its potential role in policy: perspectives from the UK Office for National Statistics. *Soc Indic Res* 2013; **114**: 73–86.

9 Goldberg DP, Williams P. *A User's Guide to the General Health Questionnaire*. NferNelson, 1988.

10 Tennant R, Hiller L, Fishwick R, Platt S, Joseph S, Weich S, et al. The Warwick-Edinburgh Mental Well-being Scale (WEMWBS): development and UK validation. *Health Qual Life Outcomes* 2007; **5**: 63.

11 Dolan P, Gudex C, Kind P, Williams A. The time trade-off method: results from a general population study. *Health Econ* 1996; **5**: 141–54.

12 NatCen Social Research and Royal Free and University College Medical School, Department of Epidemiology and Public Health. *Health Survey for England, 2010 (2nd edn)*. UK Data Archive, 2012 (http://dx.doi.org/10.5255/UKDA-SN-6986-2).

13 NatCen Social Research and University College London, Department of Epidemiology and Public Health. *Health Survey for England, 2011*. UK Data Archive, 2013 (http://dx.doi.org/10.5255/UKDA-SN-7260-1).

14 NatCen Social Research and University College London, Department of Epidemiology and Public Health. *Health Survey for England, 2012*. UK Data Archive, 2014. (http://dx.doi.org/10.5255/UKDA-SN-7480-1).

15 Hankins M. The factor structure of the twelve item General Health Questionnaire (GHQ-12): the result of negative phrasing? *Clin Pract Epidemiol Ment Health* 2008; **4**: 10.

16 Holman R, Weisscher N, Glas CAW, Dijkgraaf M, Vermeulen M, de Haan R, et al. The Academic Medical Center Linear Disability Score (ALDS) item bank: item response theory analysis in a mixed patient population. *Health Qual Life Outcomes* 2005; **3**: 83.

17 Curran PJ, Hussong AM. Integrative data analysis: the simultaneous analysis of multiple data sets. *Psychol Methods* 2009; **14**: 81–100.

18 Gibbons RD, Perraillon MC, Kim JB. Item response theory approaches to harmonization and research synthesis. *Health Serv Outcomes Res Methodol* 2014; **14**: 213–31.

19 Hawkins DM. The problem of overfitting. *J Chem Inf Comput Sci* 2004; **44**: 1–12.

20 Goldberg DP, Gater R, Sartorius N, Ustun TB, Piccinelli M, Gureje O, et al. The validity of two versions of the GHQ in the WHO study of mental illness in general health care. *Psychol Med* 1997; **27**: 191–7.

21 Smith A, Fallowfield L, Stark D, Velikova G, Jenkins V. A Rasch and confirmatory factor analysis of the General Health Questionnaire (GHQ) – 12. *Health Qual Life Outcomes* 2010; **8**: 45.

22 Wang L, Lin W. Wording effects and the dimensionality of the General Health Questionnaire (GHQ-12). *Personal Individ Differ* 2011; **50**: 1056–61.

23 Andrich D, van Schoubroeck L. The General Health Questionnaire: a psychometric analysis using latent trait theory. *Psychol Med* 1989; **19**: 469–85.

24 Castellví P, Forero C, Codony M, Vilagut G, Brugulat P, Medina A, et al. The Spanish version of the Warwick-Edinburgh Mental Well-Being Scale (WEMWBS) is valid for use in the general population. *Qual Life Res* 2014; **23**: 857–68.

25 Stewart-Brown S, Tennant A, Tennant R, Platt S, Parkinson J, Weich S. Internal construct validity of the Warwick-Edinburgh mental well-being scale (WEMWBS): a Rasch analysis using data from the Scottish health education population survey. *Health Qual Life Outcomes* 2009; **7**: 15–22.

26 Stranges S, Samaraweera PC, Taggart F, Kandala N-B, Stewart-Brown S. Major health-related behaviours and mental well-being in the general population: the Health Survey for England. *BMJ Open* 2014; **4**:e005878.

27 The EuroQol Group. EuroQol: a new facility for the measurement of health-related quality of life. *Health Policy* 1990; **16**: 199–208.

28 Thomas ML. The value of item response theory in clinical assessment: a review. *Assessment* 2011; **18**: 291–307.

29 Muthén LK, Muthén BO. *Mplus User's Guide (7th edn)*. Muhén & Muthén, 1998–2012.

30 Gibbons RD, Bock RD, Hedeker D, Weiss DJ, Segawa E, Bhaumik DK, et al. Full-information item bifactor analysis of graded response data. *Appl Psychol Meas* 2007; **31**: 4–19.

31 Reininghaus U, McCabe R, Burns T, Croudace T, Priebe S. Measuring patients' views: a bifactor model of distinct patient-reported outcomes in psychosis. *Psychol Med* 2011; **41**: 277–89.

32 Reise SP, Bonifay WE, Haviland MG. Scoring and modeling psychological measures in the presence of multidimensionality. *J Pers Assess* 2013; **95**: 129–40.

33 Schwarz G. Estimating the dimension of a model. *Ann Stat* 1978; **6**: 461–4.

34 Böhnke JR, Lutz W. Using item and test information to optimize targeted assessments of psychological distress. *Assessment* 2014; **21**: 679–93.

35 Lord FM. Quick estimates of the relative efficiency of two tests as a function of ability level. *J Educ Meas* 1974; **11**: 247–54.

36 Brown A, Croudace TJ. Scoring and estimating score precision using multidimensional IRT. In *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment* (eds SP Reise, DA Revicki): 307–33. Taylor & Francis, 2015.

37 Fliege H, Becker J, Walter O, Bjorner J, Klapp B, Rose M. Development of a computer-adaptive test for depression (D-CAT). *Qual Life Res* 2005; **14**: 2277–91.

38 Flora DB, Curran PJ. An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychol Methods* 2004; **9**: 466–91.

39 Simms LJ, Prisciandaro JJ, Krueger RF, Goldberg DP. The structure of depression, anxiety and somatic symptoms in primary care. *Psychol Med* 2012; **42**: 15–28.

40 Caspi A, Houts RM, Belsky DW, Goldman-Mellor SJ, Harrington HL, Israel S, et al. The p factor: one general psychopathology factor in the structure of psychiatric disorders? *Clin Psychol Sci* 2014; **2**: 119–37.

41 Brouwer D, Meijer RR, Zevalkink J. On the factor structure of the Beck Depression Inventory–II: G is the key. *Psychol Assess* 2013; **25**: 136–45.

42 Jovanoviæ V. Structural validity of the Mental Health Continuum-Short Form: the bifactor model of emotional, social and psychological well-being. *Personal Individ Differ* 2015; **75**: 154–9.

43 Meyer GJ, Finn SE, Eyde LD, Kay GG, Moreland KL, Dies RR, et al. Psychological testing and psychological assessment: a review of evidence and issues. *Am Psychol* 2001; **56**: 128–65.

44 Fayers PM, Hays RD. Should linking replace regression when mapping from profile-based measures to preference-based measures? *Value Health* 2014; **17**: 261–5.

45 Bhui K, Dinos S. Preventive psychiatry: a paradigm to improve population mental health and well-being. *Br J Psychiatry* 2011; **198**: 417–9.

46 Seligman MEP, Csikszentmihalyi M. Positive psychology: an introduction. *Am Psychol* 2000; **55**: 5–14.

47 Ryan RM, Deci EL. On happiness and human potentials: a review of research on hedonic and eudaimonic well-being. *Annu Rev Psychol* 2001; **52**: 141–66.

48 World Health Organization. Preamble to the Constitution of the World Health Organization as adopted by the International Health Conference, New York, 19-22 June, 1946; signed on 22 July 1946 by the representatives of 61 States (Official Records of the World Health Organization, no. 2, p. 100) and entered into force on 7 April 1948. 1946 (http://www.who.int/about.definition/en/print.html).

49 Rose G. *The Strategy of Preventive Medicine*. Oxford University Press, 1992.

50 Department of Health. *Including a Measure of Mental Wellbeing in the ONS Wellbeing Framework: Measuring National Wellbeing Technical Advisory Group Paper*. Department of Health, 2012 (http://www.ons.gov.uk/ons/guide-method/user-guidance/well-being/about-the-programme/advisory-groups/well-being-technical-advisory-group/dh-paper-for-ons-tag-meeting-on-15-april-2013.pdf).

51 Wilson M, Kellock C, Adams D, Landsberg J. *The Scottish Health Survey: Topic Report Mental Health and Wellbeing*. The Scottish Government, 2015 (http://www.gov.scot/Resource/0046/00469088.pdf).

52 Stewart-Brown S, Samaraweera PC, Taggart F, Kandala N-B, Stranges S. Socioeconomic gradients and mental health: implications for public health. *Br J Psychiatry* 2015; **206**: 461–5.

53 Huppert F, So TC. Flourishing across Europe: application of a new conceptual framework for defining well-being. *Soc Indic Res* 2013; **110**: 837–61.

54 Smith GT, McCarthy DM, Zapolski TCB. On the value of homogeneous constructs for construct validation, theory testing, and the description of psychopathology. *Psychol Assess* 2009; **21**: 272–84.

55 Reise SP, Waller NG. Item response theory and clinical measurement. *Annu Rev Clin Psychol* 2009; **5**: 27–48.

56 Wetzel E, Carstensen CH, Böhnke JR. Consistency of extreme response style and non-extreme response style across traits. *J Res Personal* 2013; **47**: 178–89.

57 Kammann R, Flett R. Affectometer 2: a scale to measure current level of happiness. *Aust J Psychol* 1983; **35**: 259–65.

58 Croudace TJ, Böhnke JR. Item bank measurement of depression: will one dimension work? *J Clin Epidemiol* 2014; **67**: 4–6.

59 Weich S, Brugha T, King M, McManus S, Bebbington P, Jenkins R, et al. Mental well-being and mental illness: findings from the Adult Psychiatric Morbidity Survey for England 2007. *Br J Psychiatry* 2011; **199**: 23–8.

Single-instrument factor analyses

This supplement reports the results for the factor analyses of item sets comprising each instrument. We assessed for each instrument (GHQ-12, WEMWBS, EQ-5D) the fit of exploratory one- and two-dimensional models to assess the potential multidimensionality of each item set. All of these analyses were performed with MPlus 7.11[29] in the estimation sample.

GHQ-12

Exploratory factor analyses ($N = 6015$; GHQ-12 data were not collected in 2011, see table 1 main document) showed that a single factor described the responses reasonably well (lowest loading $r = 0.62$; GHQ3, "playing a useful part in things"). Since previous research has focussed on a method factor for the negatively phrased items we tested this solution as well. An exploratory two-factor solution with a highly correlated second factor ($r = 0.65$; log Likelihood(LL) = -45 415, $BIC = 91\ 343$ and $BIC_{adj} = 91\ 156$) as well as a confirmatory model with one general factor and an orthogonal method factor underpinning the negative items ($LL = -45\ 377$; $BIC = 91\ 215$; $BIC_{adj} = 91\ 047$) showed better fit than the one-factor model ($LL = -46\ 368$, $BIC = 93\ 154$, $BIC_{adj} = 93\ 002$). This bifactor model is our preferred solution since it is the best-fitting as well as theoretically most appropriate of these models.

WEMWBS

A one factor solution fitted the responses reasonably well (lowest loading WEMWBS4, "interested in other people", $r = 0.60$; $N = 9569$), but the exploratory two-factor solution ($LL = -136\ 312$; $BIC = 273\ 385$; $BIC_{adj} = 273\ 121$) provided a better description of the data than one-factor solution ($LL = -137\ 929$; $BIC = 276\ 500$; $BIC_{adj} = 276\ 278$). The two factors were highly correlated ($r = 0.82$) and item 1 ("optimistic about the future"), item 4 ("interested in other people"), item 9 ("feeling close to other people"), item 12 ("been feeling loved") and item 13 ("interested in new things") loaded on this factor, capturing additional variance from items of the interest/social domain.

EQ-5D

The one-factor solution revealed evidence of one relatively low factor loading ($r = 0.50$; item 5, "anxiety and depression"; $N = 9607$) compared to those of the other four (0.85 to 0.93), indicating that the EQ-5D covers more than one latent dimension. Extending the modelling to a two-factor solution ($LL = -20\ 653$; $BIC = 41\ 480$; $BIC_{adj} = 41\ 419$) provides a slightly better description of the data than the one-factor solution ($LL = -20\ 696$; $BIC = 41\ 531$; $BIC_{adj} = 41\ 484$). These two dimensions are highly correlated ($r = 0.71$). The two dimensions separate the first four items (on one factor with loadings above $r = 0.73$) from the last item, which loads highly on the second factor ($r = 0.68$) and not at all on the first factor.

TRANSLATION INTO THE JOINT MODELLING APPROACH

Overall these results suggest that the GHQ-12 as well as the WEMWBS assess mainly one dimension. The second dimension that is extracted for these two instruments is always highly correlated with the first one and the loadings of the items from the models with only one dimension are already very high, indicating a high degree of shared variance between the items. In contrast, the EQ-5D shows the clear need for the extraction of a second dimension.

As described in the paper, the joint modelling proceeded in three steps. The first was fitting a single factor model (model 1) to the data and the second estimating a three-dimensional model (model 2). These two analyses were exploratory in nature to provide a baseline for the comparison between the models in terms of relative fit (single factor model) and to investigate the degree of separation between the three instruments (three factor model). Both did not build on any information gathered in the single scale models described above.

The third step involved fitting a confirmatory bifactor model. After fitting a model solely based on the three instruments (model 3), this step had to build on information gathered on the functioning of the individual instruments (model 4). While the structure of the GHQ-12 and EQ-5D was straightforwardly translated into the bifactor model (for the GHQ-12 a bifactor structure was already the best-fitting model; for the EQ-5D the final solution presented in online Table DS1 shows loadings of the same magnitude as identified by the single scale analysis presented in this supplement), the translation of the WEMWBS needed a slight adjustment. While fitting a bifactor model with the goal to generate independently interpretable scores for general and domain factors[32] (all WEMWBS items loading on the general factor as well as the orthogonal domain factor), we tried to estimate a second domain factor that was correlated with the first one to mirror the structure identified in the single scale analysis as closely as possible. Nevertheless, this revealed that insufficient variance was left to be explained in the WEMWBS items. Instead we added an orthogonal factor for the items of the interest/social domain to take some of the insight gathered from the single scale analysis over into the bifactor model.

Table DS1 Standardised factor loadings of the three discussed models

	Model 2: geomin exploratory			Model 1: single factor	Model 4: bifactor analysis					
	Factor 1	Factor 2	Factor 3		General	GHQ-12	WEMWBS	EQ-5D	Wording	Interest/ social
GHQ1, able to concentrate	0.34	0.52		0.58	0.63	0,42				
GHQ2, lost sleep		0.70		0.54	0.65	-0.09			**0.35**	
GHQ3 play useful part	0.36	0.40		0.54	0.59	0.44				
GHQ4, capable of making decisions	0.32	0.54		0.59	0.65	0.51				
GHQ5, under strain		0.82		0.59	0.72	-0.10			**0.49**	
GHQ6, not overcome difficulties		0.81		0.68	0.81	-0.06			0.26	
GHQ7, enjoy day to day activities	0.47	0.47		0.55	0.63	0.44				
GHQ8, able to face problems		0.64		0.65	0.72	0.39				
GHQ9, unhappy		0.84		0.73	0.87	-0.13			0.11	
GHQ10, lose confidence		0.78		0.75	0.89	-0.15			-0.11	
GHQ11, worthless person		0.70		0.72	0.85	-0.18			-0.15	
GHQ12, reasonably happy		0.61		0.65	0.75	0.24				
WEMWBS1, optimistic			0.60	0.61	0.43		**0.46**			0.10
WEMWBS2, feeling useful			0.69	0.71	0.48		**0.55**			
WEMWBS3, feeling relaxed			0.57	0.74	0.60		0.46			
WEMWBS4, interested in other people			0.72	0.56	0.31		**0.52**			**0.31**
WEMWBS5, energy to spare	0.36		0.50	0.64	0.49		**0.41**			
WEMWBS6, deal well with problems			0.68	0.79	0.58		**0.58**			
WEMWBS7, thinking clearly			0.69	0.78	0.56		**0.60**			
WEMWBS8, feel good about myself			0.73	0.88	0.67		**0.60**			

WEMWBS9, feel close to other people			0.81	0.71	0.45		**0.57**			**0.51**
WEMWBS10, feel confident			0.76	0.87	0.64		**0.63**			
WEMWBS11 make up own mind			0.66	0.70	0.48		**0.57**			
WEMWBS12, felt loved			0.69	0.62	0.43		**0.47**			**0.46**
WEMWBS13, interested in new things			0.72	0.68	0.44		**0.54**			0.17
WEMWBS14, felt cheerful			0.75	0.86	0.65		**0.58**			
EQ-5D1, mobility	0.96			0.36	0.36			**0.87**		
EQ-5D2, self-care	0.89			0.54	0.48			**0.78**		
EQ-5D3, usual activities	0.91			0.47	0.48			**0.79**		
EQ-5D4, pain	0.83			0.32	0.37			**0.77**		
EQ-5D5, anxiety/depression		0.53		0.71	0.79			0.15		
Omega estimates	--	--	--	--	$\omega = 0.98$ $\omega_h = 0.81$	$\omega = 0.96$ $\omega_s = 0.11$	$\omega = 0.95$ $\omega_s = 0.44$	$\omega = 0.94$ $\omega_s = 0.61$	--	--

GHQ-12, 12-item General Health Questionnaire; WEMWBS, Warwick-Edinburgh Mental Well-being Scale.

Item responses were recoded where necessary, so that higher values indicate a more positive response from the participant. For readability factor loadings <|0.30| are blanked for the exploratory model. Loadings >|0.30| on domain factors in the bifactor model equal (±0.1) or higher than those on the general factor are set in bold face for the three instruments (indication of multidimensionality).

Factor loadings for unidimensional model in the middle to ease comparison with the two other models; correlation between factors for geomin-rotated exploratory model: r(F1, F2) = 0.34; r(F1, F3) = 0.29; r(F2, F3) = 0.51.