Psychological Medicine (2008), **38**, 575–579. © 2008 Cambridge University Press
doi:10.1017/S003329170800281X Printed in the United Kingdom

ORIGINAL ARTICLE

A hierarchy of distress: Mokken scaling of the GHQ-30

R. Watson[1]*, I. J. Deary[2] and B. Shipley[2]

[1] *School of Nursing and Midwifery, The University of Sheffield, UK*
[2] *Department of Psychology, The University of Edinburgh, UK*

Background. Hierarchical cumulative scales are common and informative in psychology. The General Health Questionnaire (GHQ) does not appear to have been subjected to an analysis that examines the hierarchical and cumulative nature of its items. We report an analysis of data from the 30-item GHQ (GHQ-30) as part of the Health and Lifestyle Survey (HALS).

Method. Data from 6317 participants who completed the GHQ-30 as part of the HALS were analysed using the Mokken Scaling Procedure (MSP), which is a computer program that searches polychotomous data for hierarchical and cumulative scales on the basis of a range of diagnostic criteria.

Results. A final scale consisting of nine items from the GHQ-30 was obtained that, according to the criteria for a Mokken scale, was a reliable and very strong scale. The least difficult item in the scale is 'been (un)able to face up to your problems?' and the most difficult item is 'felt that life isn't worth living?'

Conclusions. Items from the GHQ-30 form a short hierarchical and cumulative scale. The majority of these items also appear in the GHQ-12. The nine-item GHQ shows better distribution properties than the GHQ-30 and compares very favourably with the GHQ-12.

Received 21 August 2007; Revised 7 December 2007; Accepted 14 December 2007; First published online 29 January 2008

Key words: General health, item response theory, Mokken scaling.

Introduction

The General Health Questionnaire (GHQ) was developed by Goldberg & Hillier (1979) as a screening device for minor, non-psychotic, psychiatric disorder, or psychological morbidity. Originally devised as a 60-item questionnaire measuring physical and psychological symptoms, the GHQ is also available in 30-, 28- and 12-item versions that focus on the psychological symptoms. For example, the 28-item questionnaire covers four main elements of distress: depression, anxiety, social impairment and insomnia (Goldberg & Hillier, 1979). The GHQ is widely used clinically and in research and has been translated into several languages (http://shop.nfer-nelson.co.uk/icat/generalhealthquestionnair; accessed 9 January 2008). Each version of the GHQ is related through having common items and, although different scoring systems can be applied, including a modified Likert-type scoring system, a total score is generated. The total score indicates the level of psychological morbidity, with higher scores indicating greater levels of morbidity, or poorer general health, with suggested thresholds, indicating psychological distress, for each version of the scale.

As far as it is possible to discern from an online search of the Web of Science using 'GHQ' and 'Mokken' as search terms and no date restrictions, the GHQ, which is strong psychometrically, has not been subjected to analysis for a hierarchy among its items, although it has been used to validate other scales that have been developed using this procedure (Barkow *et al*. 2001; Mergl *et al*. 2007). One hierarchical scale for general health has been developed but does not seem to be widely used (Moorer & Suurmeijer, 1994; Moorer *et al*. 2001). Hierarchical scales are used frequently in social, psychological, medical and nursing research (Kempen & Suurmeijer, 1991; Watson 1996; Kingshott *et al*. 1998; Ringdall *et al*. 2003), and establishing whether a scale has hierarchical properties adds a new dimension to its use other than simply using the total score obtained by summing, for example, Likert-type responses. If a scale is demonstrated to have hierarchical properties, it indicates that the items are ordered relative to one another and, by implication, ordered along the latent trait that is being measured. Therefore, although a total score from a

* Address for correspondence: Dr R. Watson, School of Nursing and Midwifery, The University of Sheffield, Sheffield S10 2TN, UK.
(Email: roger.watson@sheffield.ac.uk)

set of hierarchically ordered items indicates the extent to which the latent trait is present (or absent), just as with any item response theory, a score on any item alone in a hierarchical scale indicates the extent to which the latent trait is present. For example, using the analogy of climbing a ladder, with items representing steps on the ladder and the ladder representing the latent trait, it is obvious that you cannot reach a certain point on the ladder, say step 10, without first having reached all the steps below it; if you have only reached step 10 then you will not have reached any of the steps above it. This is the nature of a hierarchical scale.

It is surprising that the GHQ has not been analysed for a hierarchy of items because an inspection of the items of which it is composed suggests that some items, such as 'been able to concentrate on whatever you're doing?' and 'been getting out of the house as much as usual?', seem to suggest a level of psychological morbidity that is lower than, for example, 'felt that life is entirely hopeless?' and 'felt that life isn't worth living?' In the terminology of hierarchical scales, the former items seem less 'difficult' to endorse than the latter, where 'difficulty' refers to the ease with which individuals will endorse them. Presented with a list of items of varying difficulty, more people will endorse the less difficult items than the more difficult items. In relation to the description of hierarchical scales given above, it is the relative levels of difficulty of items and the extent to which pairs of items are always ordered by this difficulty that lies at the heart of establishing a hierarchical scale.

Mokken scaling

The original description of hierarchical scales for dichotomous items was provided by Guttman (Stouffer *et al*. 1950). Guttman scales are deterministic, that is they rely on people only scoring on an item and all those below it in a scale and on none of the items above it in the scale in terms of difficulty. The Guttman model has been refined to a stochastic model by Mokken (Mokken & Lewis, 1982) and this has been further refined to accommodate polychotomous items, and software, the Mokken Scaling Procedure (MSP), is available for analysis (Sijtsma *et al*. 1990). Mokken scaling, which is simply one of several item response theories, is a non-parametric method for determining whether hierarchical scales (i.e. Mokken scales) exist in an item bank. According to Hosenfeld *et al*. (1997, p. 369), 'a genuine Mokken scale meets the assumptions of both the more liberal model of monotone homogeneity and the stricter model of double monotonicity.' Monotone homogeneity means that, as the latent trait increases, so do the all of the item response curves, and double monotonicity means that item response curves do not intersect. These assumptions mean that individuals can be ordered along the latent trait being measured and also that the items show, in each individual, invariant item ordering (Sijtsma & Junker, 1996). These properties were discussed and described recently by Watson *et al*. (2007). The extent to which a set of items is scalable, in Mokken terms, is given by Loevinger's coefficient (H), which is a measure of how well the set of items meet the hierarchical criteria of Mokken scales. H can be calculated for individual items in terms of the number of times they violate hierarchical (i.e. Guttman) assumptions relative to other items and an overall H can be calculated for a set of items. Generally, $H=0.3$ is taken as the minimum value for a Mokken scale and $H \geqslant 0.4$ is considered to indicate a strong Mokken scale. Other diagnostics, indicating the reliability of the scale, the probability of obtaining the scale and the extent to which scales show monotone homogeneity and double monotonicity, are available in the MSP and these are described in the following section.

Method

Data from 9003 participants who completed the GHQ-30 as part of the Health and Lifestyle Survey (HALS) were obtained and entered onto an Excel spreadsheet. The HALS is a nationwide sample survey of all adults resident in England, Scotland and Wales. In 1984–85, 12254 addresses were randomly selected from UK Electoral Registers and, from each address, one adult aged 18 years or older was invited to participate in the study. This yielded baseline interviews with 9003 individuals aged between 18 and 99 years. The GHQ-30 was completed by the participants at home and returned by post. Each of the 30 questions in the GHQ were answered using a four-point Likert scale noting the degree to which the respondent has experienced a particular symptom ('not at all', 'no more than usual', 'rather more than usual', 'much more than usual'). Scoring was then based on the 0–0/1–1 method, where 'not at all' and 'no more than usual' are scored as 0 (symptom not experienced) and 'rather more than usual' and 'much more than usual' are scored as 1 (symptom experienced). This produces a total score ranging from 0 to 30. Endorsing at least five items is the screening threshold used to identify a probable case of psychiatric disorder. The higher the score on the GHQ-30, the higher the distress. The HALS was compared to the 1981 Census to determine whether the sample was representative of the UK general population. Although women are slightly under-represented, the HALS does provide a reasonably good representative sample (Cox *et al*. 1987).

https://doi.org/10.1017/S003329170800281X Published online by Cambridge University Press

Table 1. *Nine-term GHQ Mokken scale: overview*

Item	Mean	*H*	Labels
29	1.24	0.56	Felt that life isn't worth living?
30	1.26	0.58	Found at times you couldn't do anything because your nerves were too bad?
24	1.39	0.59	Been thinking of yourself as a worthless person?[a]
23	1.64	0.62	Been losing confidence in yourself?[a]
28	1.65	0.64	Been feeling nervous and strung-up all the time?
15	1.72	0.60	Felt you couldn't overcome your difficulties?[a]
18	1.86	0.56	Been taking things hard?
14	1.99	0.58	Felt constantly under strain?[a]
20	2.01	0.50	Been (un)able to face up to your problems?[a]

Scale: $H=0.59$; reliability $\rho=0.90$; $p=0.00011$ ($n=6317$); mean=14.77 (S.D.=4.26); skewness=1.55; kurtosis=2.93.

[a] Items in GHQ-12.

The data were transferred to SPSS for Windows version 13.0 (SPSS Inc., Chicago, IL, USA). and any subjects with missing data were removed before the data ($n=6317$) were saved as a tab-delimited file with the spreadsheet option turned off; this procedure creates a file that can be read by the MSP program software. The MSP version 5.0 for Windows run on an IBM-compatible PC was used for the analysis. Mokken scaling is an important aspect of psychometric information concerning a psychological scale. However, because it is likely to be unfamiliar, the key statistical concepts involved are now explained.

The MSP program, developed by Molenaar & Sijtsma (2000), searches polychotomous item banks for reliable, hierarchical scales. The MSP enables the analyst to diagnose for monotone homogeneity and double monotonicity among those items to ensure that items are non-intersecting, as described by Watson *et al.* (2007). The diagnostic value '*Crit*' generated by the MSP enables this diagnosis by calculating a single value from the combined *H* coefficients of the items retained in the analysis. Values of *Crit* >80 are considered to indicate violations of monotone homogeneity and double montonicity; values of zero are considered to indicate perfectly non-intersecting items and values of *Crit* <40 are considered to be the result of sampling error; therefore, it is considered acceptable to include items with *Crit* values ⩾0 or <40 (Molenaar & Sijtsma, 2000). In addition to the *Crit* value, the P(++) matrix, which shows the probability of obtaining items at certain points in the scale, can be visually inspected. The P(++) matrix should show increasing values from right to left and from top to bottom (Niemöller & van Schuur, 1983).

The reliability of the scales obtained by the MSP is obtained using a test–retest procedure analogous to Cronbach's α (Moorer & Suurmeijer, 1994), generating a statistic, ρ, that should be $\geqslant 0.7$ for a scale to be considered reliable. The probability of obtaining any scale generated is tested for taking into account the multiple steps involved in this iterative program using a Bonferroni-type method of correction (Molenaar & Sijtsma, 2000). Different start sets of items may be used to avoid capitalizing on the first pair of items identified by the MSP. Finally, summary scale statistics are generated (mean, skewness and kurtosis) to show how closely scores obtained using the final scale are normally distributed.

All 30 items were entered into the MSP and, by increasing the lower-bound *H* value incrementally in 0.05 steps from 0 to 0.50, the number of scales obtained, the number of items they contained and their reliability recorded, as recommended by Molenaar & Sijtsma (2000). This preliminary analysis continued until reliable scales with sufficiently high *H* are obtained before further analyses of homogeneity and double monotonicity are carried out.

Results

Only one reliable scale ($\rho>0.7$) containing 15 items was obtained up to $H=0.40$. Thereafter, a second reliable scale was obtained that contained only two items. Therefore, further analysis was carried out using a lower-bound *H* of 0.40. The 15 items were checked for monotone homogeneity and double montonicity and, using the diagnostic *Crit* values that were asterisked as violating these criteria down to *Crit* <40, a final scale consisting of nine items was obtained, as shown in Table 1. Using different start sets of items provided the same scale. The least difficult item in the scale is 'been (un)able to face up to

https://doi.org/10.1017/S003329170800281X Published online by Cambridge University Press

your problems?' with a mean score 2.01. The most difficult item is 'felt that life isn't worth living?' with a mean score of 1.24. The scale is very strong ($H = 0.59$) and highly reliable ($\rho = 0.90$). It should be noted that, using the scoring system applied in this study, lower mean scores relate to greater difficulty in an item even though these items relate to greater levels of distress.

Discussion

Nine items from the GHQ-30 form a reliable and strong hierarchical scale. In addition to the diagnostic criteria produced by the MSP, the scale has face validity: the items form a sensible hierarchy of difficulty. The two extremes in the scale make sense in that an early stage in psychological distress, leading to subsequent higher levels of distress, is likely to be overwhelmed by personal problems and, conversely, an extreme level of psychological distress is likely to be represented by feeling that life is no longer worth living. This extreme level of distress may be preceded by being paralysed by an inability to function and feelings of worthlessness and lower levels of distress may be associated with feeling strained, tense and sensitive to adverse events.

The authors of the GHQ have included items within the scale that form a hierarchy and, as shown by the present analysis, these could form a useful scale. The majority of the nine items (asterisked in Table 1) are included in the shortest version of the GHQ, the GHQ-12. The scores for this nine-item GHQ only approximately form a normal distribution and some further development may be necessary for the scale to be implemented in clinical studies. However, it should be noted that the skewness of scores for the parent GHQ-30 in the same sample was the same as for the nine-item GHQ (1.55) and that the kurtosis was greater (4.07). For the GHQ-12, using the same sample, the skewness was 1.36 and the kurtosis was 2.58. Therefore, the nine-item GHQ shows better distribution properties than the GHQ-30 and compares very favourably with the GHQ-12, and it has the newly discovered advantage of forming a strong hierarchical scale in this large, representative sample.

The advantage and clinical utility of such a scale are that it is shorter than all previous versions of the GHQ. However, the construct validity of the instrument should be tested further by measuring convergent validity against other measures of psychological distress. In addition, prior to clinical use, levels of psychological distress measured using this nine-item version of the GHQ would have to be established against accepted diagnostic criteria and population norms established.

Acknowledgements

Ian Deary is the recipient of a Royal Society-Wolfson Research Merit Award.

Declaration of Interest

None.

References

Barkow K, Heun R, Üstün TB, Maier W (2001). Identification of items which predict later development of depression in primary care. *European Archives of Psychiatry and Clinical Neuroscience* **251** (Suppl. 2), II21–II26.

Cox BD, Blaxter M, Buckle ALJ, Fenner NP, Golding JF, Gore M, Huppert FA, Nickson J, Roth M, Stark J, Wadsworth MEJ, Whichelow M (1987). *The Health and Lifestyle Survey: A Preliminary Report*. Health Promotion Trust: London.

Goldberg DP, Hillier VF (1979). A scaled version of the General Health Questionnaire. *Psychological Medicine* **9**, 139–145.

Hosenfeld B, van den Boom DC, Resing WCM (1997). Constructing geometric analogies for the longitudinal testing of elementary school children. *Journal of Educational Measurement* **34**, 367–372.

Kempen GIJM, Suurmeijer ThPBM (1991). Factors influencing professional home care utilization among the elderly. *Social Science and Medicine* **32**, 77–81.

Kingsott R, Douglas N, Deary I (1998). Mokken scaling of the Epworth Sleepiness Scale items in patients with sleep apnoea/hypopnoea syndrome. *Journal of Sleep Research* **7**, 293–294.

Mergl R, Seidscheck I, Allgaier A-K, Möller H-J, Hegerl U, Henkel V (2007). Depressive, anxiety, and somatoform disorders in primary care: prevalence and recognition. *Depression and Anxiety* **24**, 185–195.

Mokken RJ, Lewis C (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement* **6**, 417–430.

Molenaar IW, Sijtsma K (2000). *MSP5 for Windows*. Groningen: iec ProGAMMA.

Moorer P, Suurmeijer TPBM (1994). A study of the unidimensionality and cumulativeness of the MOS Short-Form General Health Survey. *Psychological Reports* **74**, 467–470.

Moorer P, Suurmeijer ThPBM, Foets M, Molenaar IW (2001). Psychometric properties of the RAND-36 among three chronic diseases (multiple sclerosis, rheumatic diseases and COPD) in the Netherlands. *Quality of Life Research* **10**, 637–645.

Niemöller K, van Schuur W (1983). Stochastic models for unidimensional scaling: Mokken and Rasch. In *Data Analysis and the Social Sciences* (ed. D. McKay, N. Schofield and P. Whiteley), pp. 120–170. Francis Pinter: London.

Ringdall GI, Jordhøy MS, Kaasa S (2003). Measuring quality

https://doi.org/10.1017/S003329170800281X Published online by Cambridge University Press

of palliative care: psychometric properties of the FAMCARE Scale. *Quality of Life Research* **12**, 167–176.

Sijtsma K, Debets P, Molenaar IW (1990). Mokken scale analysis for polychotomous items: theory, a computer program and an empirical application. *Quality and Quantity* **24**, 173–188.

Sijtsma K, Junker BW (1996). A survey of theory and methods of invariant item ordering. *British Journal of Mathematical Statistics and Psychology* **49**, 79–105.

Stouffer SA, Guttman L, Suchman EA, Lazarsfeld PF, Star SA, Clausen JA (1950). *Measurement and Prediction*, vol. 4. Princeton University Press: Princeton, NJ.

Watson R (1996). The Mokken scaling procedure (MSP) applied to the measurement of feeding difficulty in elderly people with dementia. *International Journal of Nursing Studies* **33**, 385–393.

Watson R, Deary IJ, Austin E (2007). Are personality trait items reliably more or less 'difficult'? Mokken scaling of the NEO-FFI. *Personality and Individual Differences* **43**, 1460–1469.

https://doi.org/10.1017/S003329170800281X Published online by Cambridge University Press