Comparison of the General Health Questionnaire and the Hospital Anxiety and Depression Scale.

G Lewis and S Wessely

BJP 1990, 157:860-864.
Access the most recent version at DOI: 10.1192/bjp.157.6.860

British Journal of Psychiatry (1990), **157**, 860–864

Comparison of the General Health Questionnaire and the Hospital Anxiety and Depression Scale

GLYN LEWIS and SIMON WESSELY

The specificity and sensitivity of the HAD, 12-item GHQ and CIS were calculated by comparing the scores of dermatological patients on these tests with a criterion measure of disorder. Since psychiatry, along with many other branches of medicine, does not have an error-free criterion, it was assumed that the criterion was an underlying latent construct which was measured by all of the tests and could be derived by factor analysis from the scores on them. No differences were found between the two questionnaires (HAD and GHQ) in their ability to detect cases of minor psychiatric disorder although they were somewhat less reliable than the CIS.

It is commonplace to study the validity of simple tests by comparing them with a 'gold standard' or criterion and thus calculating the sensitivity and specificity. These validity indices thus give an estimate of the accuracy with which a test measures the criterion.

Unfortunately psychiatrists, along with many of their medical colleagues, do not have a criterion measure of disorder. All the currently used standardised interviews, which are traditionally regarded as the criterion, are not free of error and show modest repeatability. Indeed, it is ironic that the sensitivity and specificity of the more repeatable General Health Questionnaire (GHQ; Goldberg, 1972) is calculated in relation to the less repeatable standardised interviews. When the criterion shows error, the sensitivity and specificity of the test will apparently be reduced. Furthermore, conventional statistics only estimate sampling error and do not account for variation resulting from error in the criterion.

The main aim of this paper is to use factor-analytic models as a form of validity testing in addition to the traditional method of comparing a questionnaire with a standardised interview. Such models assume that the criterion is an underlying construct (in this case, overall severity of minor psychiatric illness) which is measured by the questionnaires and interviews and can be derived by factor analysis from the scores on them.

Screening for psychiatric disorder in the physically ill

There has been much debate about the validity of using self-report questionnaires in the physically ill (Kirk & Saunders, 1979; Mayou & Hawton, 1986; Goldberg & Williams, 1988) and other authors have suggested that the GHQ score used to define a 'case' should be raised in those with neurological conditions (Bridges & Goldberg, 1986) or angina (Vasquez-Barquero *et al*, 1985). This is usually attributed to the overlap between the somatic symptoms of psychological disorder and those of physical disease. False positives are also explained by questions in the GHQ assessing social functioning, for instance "Have you recently been getting out of the house as much as usual?" The influence of physical illness on assessments of psychiatric disorder is a potential source of bias in much psychiatric research.

Structured interviews partly rely upon the interviewer (usually a psychiatrist) making judgements based on clinical impression. This could lead to an underestimate of the prevalence of psychiatric disorder in the physically ill. Firstly, symptoms (e.g. fatigue, insomnia) may be attributed to physical illness, or omitted from the score through caution, although in fact they are caused by psychological illness. Secondly, the emotional consequences of physical illness could be viewed as a normal or understandable response to a life-threatening diagnosis rather than a psychiatric disorder. Both these factors would reduce the apparent prevalence of psychiatric disorder in the physically ill and lead to an increase in the proportion of false positives. Again, the absence of a criterion prevents an easy solution to this problem.

Zigmond & Snaith (1983) devised the Hospital Anxiety and Depression Scale (HAD) specifically for detecting anxiety and depression in the physically ill, and it is therefore of interest to compare the GHQ with the score derived from combining the two scales of the HAD. The main differences between the GHQ

and the HAD are that the latter asks only about psychological symptoms and does not have the 'same as usual' option, a response which scores zero in the GHQ.

The HAD has been compared with clinical rating scales (Zigmond & Snaith, 1983; Aylard *et al*, 1987) and standardised interviews (Barczak *et al*, 1988; Andrews *et al*, 1988) but only two papers have compared the HAD with the GHQ. Wilkinson & Barczak (1988) compared the HAD with the GHQ-28 in a general-practice sample, using the structured clinical interview for DSM-III as the criterion. Unfortunately there was no indication whether the slight superiority of the HAD could be attributed to chance. Aylard *et al* (1987) also compared the HAD with the GHQ-28 but like Zigmond & Snaith (1983) they did not use a single threshold to define a case and so their studies produced inflated values for sensitivity and specificity by excluding doubtful cases.

Using factor-analytic models, this study compared the HAD, the 12-item GHQ and the Clinical Interview Schedule (CIS; Goldberg *et al*, 1970), in their ability to detect cases of minor psychiatric disorder in a sample of new referrals in a hospital setting. Using this method the two self-report questionnaires could be compared with each other and with the CIS.

The study was conducted in a dermatology clinic and dermatological conditions do not usually lead to impaired social functioning. There is little empirical work on patients with dermatological disease and although these disorders are reputed to frequently arise psychogenically there is no quantitative evidence that this is so. Wessely & Lewis (1989) concluded that there is little evidence to suggest either that dermatology patients have an excessive prevalence of psychiatric conditions or that there is an obvious link between particular skin disorders and psychiatric morbidity. The 12-item GHQ was chosen partly because it has a similar number of items to the HAD (14 items) and also because it contains no somatic symptoms. Goldberg & Williams (1988) concluded, on the basis of an extensive review and meta-analysis, that the validity of the GHQ-12 is comparable with that of the longer versions of the GHQ.

Method

On each day selected for study, a random sample of all new patients attending the Dermatology Out-patient Clinic of King's College Hospital were invited to take part in the study. Each subject was asked to complete the HAD and 12-item GHQ and to be interviewed by one of the authors using the CIS. This was part of a larger study where other information was collected and these results have been presented elsewhere (Wessely & Lewis, 1989).

A 'case' was defined if the subject scored 13 or more on the total weighted score of the CIS (Goldberg *et al*, 1970). Relative operating characteristic (ROC) curves were fitted using the ROCFIT program (Metz *et al*, 1984).

The GHQ was scored using the traditional method (0, 0, 1, 1; Goldberg, 1972) except for analyses in which factor analysis was used. In these cases the Likert (0, 1, 2, 3) method of scoring was adopted. This method leads to a distribution of scores which more closely approximates the normal distribution.

Use of factor-analytic measurement models

The sensitivity, specificity and kappa values of a self-report questionnaire, traditionally assessed by comparison of the questionnaire with a standardised interview (the criterion), may be reduced as a result of error in the criterion as well as that in the test itself. This should not affect the comparison of two tests except that conventional statistics only estimate sampling error and so apparently statistically significant differences between tests should be interpreted cautiously. One way of attempting to circumvent this problem is by use of factor-analytic measurement models (Allen & Yen, 1970; Joreskog, 1971). The model assumes that all three psychiatric measures used here are measuring the same underlying construct which can be represented by a single factor derived by maximum-likelihood factor analysis from the scores on the three tests. All three measures can then be compared with the factor which is used as the criterion. In this study the LISREL VI program was used (Joreskog & Sorbom, 1974).

Two analyses can be performed using this method. Firstly, the reliabilities of the scales can be calculated using LISREL VI. Assuming that all variation apart from the factor is random error, LISREL can calculate the variance of this random-error term. From classical test theory, the reliability is given by the variance of the true score divided by the total variance (Carmines & Zeller, 1979). The variance of the factor is used as the estimate of the true variance and the total variance is therefore calculated by adding the factor variance to the variance of the random-error term calculated using LISREL. This estimate of reliability can be conceived as the accuracy with which each test measures the factor-derived construct.

The second analysis performed determines a threshold score on the factor and this is used instead of the CIS to define a 'case'. In this study the threshold on the factor was determined using ROC analysis (Mari & Williams, 1985). All three measures can then be compared with the factor-derived case definition.

In summary, the assumptions underlying the measurement model are as follows:

(a) the two questionnaires and the total weighted score of the CIS are measuring the same psychological construct
(b) this construct can be represented by a single factor derived by maximum-likelihood factor analysis from the three measures
(c) each measure can be represented by a linear combination of the factor and a random-error term;

TABLE I

Comparison of: (a) the HAD with the GHQ when the CIS is used as the criterion; and (b) the HAD, GHQ and CIS when the presumed underlying factor score is used to define a case

		Se%	*Sp%*	*Bias*	*Kappa*	*s.e.*	*ROC area*	*s.d.*
(a)	HAD	72.3	77.1	1.06	0.490	0.082	0.88	0.033
	GHQ	78.7	77.5	1.13	0.550	0.078	0.84	0.046
(b)	CIS	89.8	95.6	0.96	0.859	0.048	0.98	0.012
	GHQ	85.7	85.3	1.06	0.704	0.066	0.91	0.032
	HAD	77.6	82.4	1.02	0.597	0.075	0.92	0.026

1. Se% = sensitivity.
2. Sp% = specificity.

errors in the different measures are assumed to be uncorrelated

(d) the test scores are assumed to be normally distributed.

Results

The sample consisted of 173 subjects of whom 117 (68%) completed the HAD, GHQ and CIS. Some data were collected on almost all the 56 non-responders. They did not complete all the measures, either because the experimenters could not contact them or they claimed they could not afford the time. The average age of the non-responders ($n = 53$) was 45 years and of the responders 40 years ($t = 0.3$, d.f. 166, $P > 0.5$). Of the 56 non-responders, 55% were women while 60% of the responders were female ($\chi^2 = 0.35$, $P > 0.3$). Of the non-responders, 35% (9/26) were 'cases' defined by the 1/2 cut-off on the GHQ-12, while 44% of the responders were cases using the same criterion ($\chi^2 = 0.84$, $P > 0.3$). These results indicate that the group on whom complete data were collected was fairly representative of the original random sample.

The CIS suggested that the prevalence of minor psychiatric disorder was 40.2% (47/117; 95% CI 35.7–44.7%).

The mean score on the HAD anxiety scale (HADA) was 6.9 (s.d. = 4.23) and on the HAD depression scale (HADD) was 3.9 (s.d. = 3.41). The correlation between the CIS score and the HAD score (the sum of HADA and HADD) was 0.79, between the CIS and the GHQ was 0.76 and between the GHQ and HAD was 0.75.

Comparison of questionnaires using the CIS

Using the CIS-defined case as the criterion, the GHQ and HAD were compared. The recommended cut-off score for the 12-item GHQ is 1/2 (i.e. two or more is a case) but there is no accepted score for the HAD, although Wilkinson & Barczak (1988) and Andrews *et al* (1988) have suggested a 7/8 threshold. ROC analysis was used to define the best HAD cut-off (Mari & Williams, 1985), and in this sample it was 10/11. Using the sample to define the cut-off in this way tends to enhance the sensitivity and specificity of the HAD.

The sensitivities and specificities of the two questionnaires compared with the CIS are given in the first two columns of Table I(a), and the test bias (the ratio of test:criterion prevalence) is shown in the third column. The areas under the ROC curves (sixth column) give a measure of the ability of a test to distinguish a case from a non-case and an area of 1.0 indicates perfect agreement. The two ROC areas were not significantly different ($z = 0.87$; $P = 0.4$) using the method of Hanley & McNeil (1983).

The kappa statistic (Fleiss, 1981) is a measure of agreement which corrects for chance. It was also calculated with standard errors for both questionnaires (Table I; fourth and fifth columns). The kappas were compared using a resampling or 'bootstrap' method. Based upon 250 bootstrap samples, the mean difference between the kappa for the HAD and the GHQ was 0.074 and the standard error was 0.089, indicating there was not a significant difference ($P = 0.4$).

The results for the GHQ show a slightly superior sensitivity and kappa, although the ROC area for the GHQ is somewhat smaller. This apparent discrepancy is probably because the sensitivity, specificity and kappa are calculated using a single threshold while the ROC area is calculated using several thresholds. In this instance the difference between the ROC areas is not statistically significant, but, in general, the results of the ROC area calculated with a curve-fitting program is probably a better estimate of agreement between test and criterion than indices calculated using a single threshold.

In conclusion, it appears there is no substantial difference between the two questionnaires when they are compared with the CIS.

Comparison of questionnaires using measurement models

The reliabilities calculated using LISREL were 0.83 for the CIS, 0.72 for the GHQ and 0.74 for the HAD.

Using the factor scores to define a case led to the results shown in Table I(b). Again, there appeared to be few differences between the GHQ and the HAD, and the difference in ROC areas between the HAD and GHQ was not significant ($z = 0.13$, $P = 0.9$; Hanley & McNeil, 1983). However, the levels of agreement with the factor-derived criterion was lower for the two questionnaires than for the CIS, and there was a significant difference between the ROC areas for the GHQ and CIS ($z = 2.3$, $P = 0.02$).

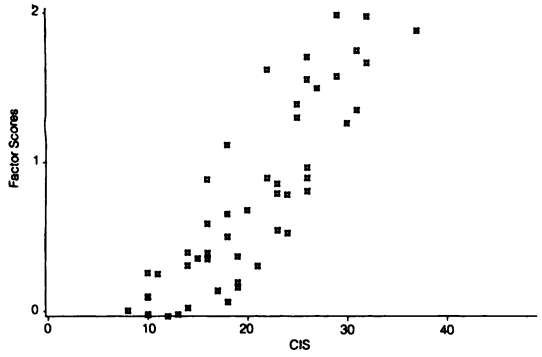

FIG. 1 Scores on the CIS for each individual plotted against the factor derived from the scores on the CIS, GHQ and HAD.

The total weighted scores on the CIS for each individual are plotted against the factor scores in Fig. 1. This illustrates that using the model, the CIS appears to be closely related to the construct represented by the factor, but is not free of error.

Discussion

In all the comparisons here, there were few differences between the HAD and the 12-item GHQ and they performed very similarly in this sample of dermatology patients. Comparison of the GHQ with other questionnaires, similar to the HAD in some respects, have found no differences in the ability of the various questionnaires to detect cases of minor psychiatric disorder (Goldberg *et al*, 1976; Mari & Williams, 1985).

Wilkinson & Barczak (1988) have argued that the HAD has a 'consistent threshold score' of 7/8. The difference with the present results (threshold 10/11) may reflect the different case definitions in the two standardised interviews, but it seems wise to avoid a fixed threshold for the HAD before it has been more widely studied.

In addition to the traditional approach to validity using a standardised interview as the criterion, factor-analytic measurement models were also used in an attempt to derive a better estimate of the underlying mental illness construct. The assumptions of the model, outlined earlier, are parsimonious and reasonable in the light of current psychiatric research into minor psychiatric morbidity. This was illustrated in Fig. 1. The analytic method seems to provide useful additional information to that obtained using the more established analyses.

In comparison with the differences between the two questionnaires, which were tiny, the measurement model suggested a small but definite advantage in the reliability of the CIS interview. This is reassuring, but not surprising as the CIS takes about ten times as long and contains many more questions than the GHQ or HAD. However, the difference in reliability is relatively modest and this questions the routinely made assumption that a standardised interview is free of error. It will be of interest to see whether interactive computerised assessments (Lewis *et al*, 1988), which can be longer and more thorough than pen and paper questionnaires, will also be more reliable.

The measurement model permits the calculation of the reliability of the observed variables in measuring the underlying construct. Reliability is a somewhat ambiguous term and the term repeatability has been recommended when the agreement between replicate measurements is estimated (Rose & Barker, 1978*a*). The circumstance here is quite different. The reliabilities are estimating the accuracy with which the tests are measuring the underlying construct, the presumed criterion. This method is therefore a form of validity testing, where the criterion is a continuous measure derived by factor analysis. Sensitivity and specificity measure the accuracy with which a test measures a categorical criterion. It has been persuasively argued that disease is best conceived as a continuum (Rose & Barker, 1978*b*) and this view seems especially appropriate for minor psychiatric disorders.

Conclusions

Measurement models have been used to try to overcome the problems of using standardised interviews as an error-free criterion. It seems that reasonable assumptions are needed for these models, and their use could provide additional insights into psychiatric measurement.

The results indicate that there were few differences between the HAD and GHQ in their ability to detect cases of minor psychiatric disorder in this setting. The standardised interview seemed to be somewhat more accurate than the self-report questionnaires.

Acknowledgements

The General Practice Research Unit is funded by the DoH and is under the direction of Professor Michael Shepherd. We would like to thank Dr Graham Dunn who provided statistical advice, an introduction to the use of factor-analytic measurement models and who performed the calculations using the resampling techniques. Dr Paul Williams also provided invaluable discussion about these issues. Our thanks also go to Drs Du Vivier, Pembroke and colleagues for their assistance in providing facilities and access to their patients. SW is supported by the Wellcome Trust.

References

ALLEN, M. J. & YEN, W. M. (1979) *Introduction to Measurement Theory*. Monterey, California: Brooks-Cole.

ANDREWS, H., BARCZAK, P. & ALLAN, R. (1988) Psychiatric illness in patients with inflammatory bowel disease. *Gut*, **12**, 1600–1604.

AYLARD, P. R., GOODING, J. H., MCKENNA, P. J., *et al* (1987) A validation study of three anxiety and depression self-assessment scales. *Journal of Psychosomatic Research*, **31**, 261–268.

BARCZAK, P., KANE, N., ANDREWS, S., *et al* (1988) Patterns of psychiatric morbidity in a genito-urinary clinic: a validation of the Hospital Anxiety and Depression scale. *British Journal of Psychiatry*, **152**, 698–700.

BRIDGES, K. W. & GOLDBERG, D. P. (1986) The validation of the GHQ-28 and the use of the MMSE in neurological in-patients. *British Journal of Psychiatry*, **148**, 548–553.

CARMINES, E. G. & ZELLER, R. A. (1979) *Reliability and Validity Assessment*. Beverly Hills: SAGE.

FLEISS, J. L. (1981) *Statistical Methods for Rates and Proportions*. New York: Wiley & Sons.

GOLDBERG, D. P. (1972) *The Detection of Psychiatric Illness by Questionnaire*. London: Oxford University Press.

—— & WILLIAMS, P. (1988) *A User's Guide to the General Health Questionnaire*. Windsor: NFER-NELSON.

——, COOPER, B., EASTWOOD, M. R., *et al* (1970) A standardised psychiatric interview for use in community surveys. *British Journal of Preventive and Social Medicine*, **24**, 18–23.

——, RICKELS, K., DOWNING, R., *et al* (1976) A comparison of two psychiatric screening tests. *British Journal of Psychiatry*, **129**, 61–67.

HANLEY, J. A. & MCNEIL, B. J. (1983) A method of comparing areas under Receiver Operating Characteristic curves derived from the same cases. *Radiology*, **148**, 839–843.

JORESKOG, K. G. (1971) Statistical analysis of sets of congeneric sets. *Psychometrika*, **36**, 109–135.

—— & SORBOM, D. (1984) *LISREL VI: Analysis of Linear Structural Relationships by the Method of Maximum Likelihood*. Indiana: Scientific Software Ltd.

KIRK, C. A. & SAUNDERS, M. (1979) Psychiatric illness in a neurological out-patient department in North East England. *Acta Psychiatrica Scandinavica*, **60**, 427–437.

LEWIS, G., PELOSI, A. J., GLOVER, E. *et al* (1988) The development of a computerised assessment for minor psychiatric disorder. *Psychological Medicine*, **18**, 737–745.

MARI, J. de J. & WILLIAMS, P. (1985) A comparison of the validity of two psychiatric screening questionnaires (GHQ-12 and SRQ-20) in Brazil, using Relative Operating Characteristic (ROC) analysis. *Psychological Medicine*, **15**, 651–659.

MAYOU, R. & HAWTON, K. (1986) Psychiatric disorder in the general hospital. *British Journal of Psychiatry*, **149**, 172–190.

METZ, C. E., WANG, P.-L. & KRONMAN, H. B. (1984) University of Chicago: ROCFIT, Department of Radiology and the Franklin McLean Memorial Hospital Research Institute.

ROSE, G. & BARKER, D. J. P. (1978*a*) Repeatability and validity. *British Medical Journal*, *ii*, 1070–1071.

—— & —— (1978*b*) What is a case? Dichotomy or continuum? *British Medical Journal*, *ii*, 873–874.

VASQUEZ-BARQUERO, J. L., ACERO, J. A. P., MARTIN, C. P., *et al* (1985) The psychiatric correlates of coronary pathology: validity of the GHQ-60 as a screening instrument. *Psychological Medicine*, **15**, 589–596.

WESSELY, S. & LEWIS, G. H. (1989) The classification of psychiatric morbidity in attenders at a dermatology clinic. *British Journal of Psychiatry*, **155**, 686–691.

WILKINSON, M. J. B. & BARCZAK, P. (1988) Psychiatric screening in general practice: comparison of the general health questionnaire and the hospital and anxiety depression scale. *Journal of the Royal College of General Practitioners*, **38**, 311–313.

ZIGMOND, A. S. & SNAITH, R. P. (1983) The Hospital Anxiety and Depression Scale. *Acta Psychiatrica Scandinavica*, **67**, 361–370.

*Glyn Lewis, MSc, MRCPsych, *Lecturer, General Practice Research Unit, Institute of Psychiatry, De Crespigny Park, London SE5 8AF*; Simon Wessely, MSc, MRCP, MRCPsych, *Senior Registrar, National Hospital for Nervous Diseases, Queen Square, London WC1*

*Correspondence