British Journal of Psychiatry (1985), **146,** 55–61

Chronicity and the General Health Questionnaire

M. E. GOODCHILD and PAUL DUNCAN-JONES

Summary: We propose a new scoring for Goldberg's (1972) General Health Questionnaire. We argue that the response 'no more than usual', to an item describing pathology, should be treated as an indicator of chronic illness rather than of good health, and we score these responses accordingly. We give evidence that this set of responses is associated with other measures of neurotic illness, and that the revised scoring provides a better prediction of caseness than the conventional scoring. The revised scoring is more strongly associated with trait neuroticism, and is more stable in repeated measurement. It is recommended in preference to the conventional scoring for most research and epidemiological purposes.

The General Health Questionniare (GHQ) is a well-known and extensively validated screening questionnaire for functional psychiatric illness (Goldberg, 1972, 1978). It has been tested and validated in a number of cultures and languages (e.g. Harding, 1976; Munoz *et al*, 1978; Chan & Chan, 1983). However, it has sometimes been criticised on the grounds that it may fail to detect chronic neurotic illnesses (Finlay-Jones & Murphy, 1979). It occurred to us that this apparent weakness in an otherwise satisfactory instrument might be due to the wording of the answer categories for some of the items, in conjunction with the accepted scoring convention. Specifically, the response 'no more than usual' to the item describing a symptom or pathological condition (such as 'been having restless, disturbed nights' or 'been feeling unhappy or depressed'), might indicate chronic illness, although such a response would conventionally be scored as healthy.

In the GHQ manual, Goldberg (1978) states that "if the questionnaire is to be useful in consulting settings it must focus on breaks in normal function rather than on lifetime traits." Whether this is true for all consulting settings is debatable, and we do not believe it is correct for epidemiological work. A chronic condition usually indicates heightened vulnerability; it implies a higher probability that a person will have an acute condition on any given occasion (Duncan-Jones, 1981). Thus, we take issue with Goldberg's (1972) comment that the 'as usual' response "merely indicates normality". It is true that the individual normally feels this way, but the response does not indicate normality in the psychiatric sense. Individuals giving this response may, we believe, be just as ill as those responding 'more than usual', though with a chronic, rather than acute or episodic condition. We have identified a sub-set of items in the 30-item GHQ where the 'no more than usual' response may alternatively indicate a chronic illness. We therefore assessed a scoring system which treats these responses as indicators of illness rather than of health. If these responses do indeed indicate chronicity rather than good health, the GHQ will provide a better index of 'present state' if they are taken into account. Below, we present details of various relevant sub-scores, compare the validity of the conventional and the proposed revised scoring in relation to standardised psychiatric interview, examine the scores longitudinally, and investigate construct validity. We also present norms, specificity, and sensitivity for the revised score.

The present investigation is largely descriptive and exploratory, but some hypotheses can be stated. These are: (1) that for those items where the 'as usual' response may indicate chronicity, the use of this response will correlate positively with other measures of psychiatric morbidity, whereas the 'as usual' response for the other items will correlate negatively with the same morbidity measures: (2) that the revised scoring will correlate more strongly with other measures of morbidity than the conventional scoring: and (3) that it will also correlate more strongly with trait neuroticism.

Method

We used data from a 30-item version of the GHQ, which corresponded to the standard set of 30 items (Goldberg, 1972, Appendix 6), except that the item 'been taking things hard' was omitted, since pilot work had shown that it was not well understood in Australia. The item 'tended to lose interest in your day-to-day activities' was substituted. The authors and five colleagues each independently identified items where the 'as usual' response (e.g. 'same as usual', 'no more than usual', 'about the same') could well indicate chronic illness. There

https://doi.org/10.1192/bjp.146.1.55 Published online by Cambridge University Press

TABLE I
"Negative" and "Positive" item sets

'Negative' set	'Positive' set
Lost much sleep over worry	Been able to concentrate on whatever you're doing
Been having restless, disturbed nights	Been managing to keep yourself busy and occupied
Felt constantly under a strain	Been getting out of the house as much as usual
Felt that you couldn't overcome your difficulties	Been managing as well as most people would in your shoes
Been finding life a struggle all the time	Felt on the whole you were doing things well
Been getting scared and panicky for no good reason	Been satisfied with the way you've carried out your task
Found everything getting on top of you	Been able to feel warmth and affection to those near you
Been feeling unhappy and depressed	Been finding it easy to get on with people
Been losing confidence in yourself	Spent much time chatting with people
Been thinking of yourself as a worthless person	Felt that you were playing a useful part in things
Felt that life is entirely hopeless	Felt capable of making decisions about things
Tended to lose interest in your day-to-day activities	Been able to enjoy your normal day-to-day activities
Been feeling nervous and strung-up all the time	Been able to face up to your problems
Felt that life isn't worth living	Been feeling hopeful about your own future
Found at times you could not do anything because your nerves were too bad	Been feeling reasonably happy all things considered

was very high concordance between six of the sets of ratings, suggesting that in those 15 items describing pathology, where the most healthy response was 'not at all', the response 'no more than usual' should be considered as a potential indicator of chronic difficulty or illness. The other 15 items all refer to positive or healthy states or activities. We shall refer to the first group of 15 items as 'Set N' (negative) and the remaining 15 items as 'Set P' (positive). The two sets of items are shown in Table I.

The four response categories for each item are usually coded either 0–3 or 1–4. For clarity, we shall refer to them as categories or codes 1–4, with 1 representing the most 'healthy', and 4 the most 'ill' response. Conventionally, each item is scored by setting codes 1 and 2 equal to 0 and codes 3 and 4 equal to 1, though Likert scoring (0–3) has also been used. In the revised scoring below, code 2 ('no more than usual') is also scored 1 (i.e. counted as 'ill') in set N, but not in set P.

		Likert	Conventional	Revised (N items)
Not at all	Code 1	0	0	0
No more than usual	Code 2	1	0	1
Rather more than usual	Code 3	2	1	1
Much more than usual	Code 4	3	1	1

Within these sets, we investigate the performance of scores for the number of items coded 1, the number coded 2, and the number coded 3 or 4. Codes 3 and 4 will be grouped together, since code 4 is quite rare. We therefore define:

N1 =the number of times a respondent uses code 1 for the N items;
N2 =the number of times a respondent uses code 2 for these items;
N34 =the number of times he uses code 3 or code 4 for the same items;
P1 =the number of times a respondent uses code 1 for the P items;
P2 =the number of times he uses code 2 for these items; and
P34 =the number of times he uses code 3 or code 4 for these items.

We also define:

N234 = N2 + N34;
GHQ = N34 + P34; and
CGHQ = GHQ + N2

i.e. N234 is the total chronic and acute score based only on the N items, GHQ is the conventional scoring, and CGHQ is our proposed revised scoring.

Our data base, which has been fully described by Henderson *et al* (1981), represents a sample of the general adult population of Canberra, drawn from the electoral role. We use three sub-sets of the data.

The first is the intial sample of 756, for whom we present GHQ data obtained at first interview. As a result of missing data, the effective sample is 753. We also use this sample to examine socio-demographic breakdowns and relationships with Zung's (1965) Self-rating Depression Scale (SDS) and another brief neurotic symptom index, the 4–NS (Henderson *et al*, 1981).

The second is a sub-sample of 157, for whom we additionally have data from the Present State Examination (PSE) (Wing *et al,* 1974). This sample was selected from the initial sample with stratification by GHQ score, and unequal probability of selection, as described by Henderson *et al* (1981). Using appropriate weighting, this sample yields estimates for GHQ/PSE data that have 25–40% greater precision than a simple random sample of the same size. We have therefore taken the effective sample size as 190 for assessment of statistical significance.

The third is a longitudinal panel sub-sample, who were interviewed four times, at four-month intervals. We have GHQ data, and also life event distress scores (Henderson *et al,* 1981) at all time-points, the DSSI state anxiety/state depression scale (Foulds, 1976) collected at waves 3 and 4, and the EPI N score (trait neuroticism, Eysenck & Eysenck, 1964) from waves 2 and 4. The two neuroticism scores are averaged to give a more reliable measure of the trait. After discarding one respondent with missing EPI data, the number of observations in this set is 230. Throughout, we have used all GHQs for which valid data were available for at least 28 of the 30 items.

Results

The initial sample

Table II presents a summary comparison of Set N and Set P. The average item distribution is markedly different in the two sets, the chronic 'as usual' responses (code 2 in the N set) being much rarer than the 'as usual' responses in the P set. Table III shows that the CGHQ has a much higher mean than the GHQ, a rather larger standard deviation, and very much less skew. In some applications, the large number of zero scores in the GHQ causes problems. ('floor effect'); with the CGHQ , however, this problem disappears.

The two scoring systems both have excellent reliability (alpha). The distributions of the two scores are shown in Figure 1 (a & b). Table IV shows Spearman rank correlations of the sub-scores with two other questionnaire measures of psychiatric illness. Both measures are positively correlated with N2 and negatively correlated with P2, in line with our hypothesis. Anomalously, P1 is uncorrelated with these measures, though it represents the (supposedly) healthiest responses to the positive items—'better than usual'. The Table also shows the CGHQ to be somewhat more highly correlated with the other measures than the GHQ. Correlations of the sub-scores with self-rated general health and with an index of minor physical symptoms show the same pattern more weakly.

The GHQ and CGHQ were compared in relationship to socio-demographic variables. In this data-set, the GHQ has only a very slight (but significant) sex difference; this difference vanishes for the CGHQ. The GHQ has a very slight linear relationship to age, whereas the CGHQ is unrelated to age. Neither score has any significant relationship to educational level.

The PSE data

Using the PSE Index of Definition (Wing, 1976) as criterion, logit regressions were fitted, with GHQ and CGHQ as predictors. To define 'caseness', the Index of Definition was dichotomised between ID4 (specific symptoms) and ID5 (threshold). Using Efron's (1978) entropy coefficient as a measure of explained variance, the GHQ accounts for 26.4% of the variance in caseness, and the CGHQ for 34.9%. The CGHQ performs significantly better than the GHQ (P <.01). We also compare the two scoring systems in terms of 'sensitivity' (the proportion of cases that fall above a cut-off value on the screening score) and 'specificity' (the proportion of non-cases falling at or below the same cut-off) (Yerushalmy, 1947; Lilienfeld, 1976). Smoothed estimates of sensitivity and specificity were obtained for all possible score values, using the estimated numbers of cases at each score level from the logit regression to give more robust, model-based estimates. We shall give estimates for the score-cutting points where the two indices are most nearly equal. For the GHQ (cutting point 4/5), sensitivity is 73.5%, and specificity 76.4%. The CGHQ (12/13) gives a sensitivity of 84.0% and specificity of 80.2%. Spearman correlations of the CGHQ (GHQ) with the PSE scores were: total score 0.58 (0.52); specific neurotic reaction 0.43 (0.31); and non-specific neurosis 0.63 (0.61). It appears that the proposed modifica-

TABLE II

Averaged % distribution of item responses (n = 753)

	"N" set	"P" set
Code 1	56.2	20.2
Code 2	32.6	70.0
Code 34	11.2	9.8
	100.00	100.00

TABLE III

Summary statistics for GHQ and CGHQ (n = 753)

	GHQ	CGHQ
Mean	3.13	7.81
Standard Deviation	4.47	5.65
Skewness	2.19	.81
Median	1.44	6.67
% Zero Score	36.0	5.0
Alpha	.90	.89

TABLE IV

Rank correlations of sub-scores with other measures of psychiatric illness (n = 753)

	4–NS	Zung SDS
N1	–.49	–.56
N2	.25	.36
N34	.50	.44
P1	.03	–.06
P2	–.27	–.27
P34	.35	.46
GHQ	.46	.50
CGHQ	.50	.59

https://doi.org/10.1192/bjp.146.1.55 Published online by Cambridge University Press

TABLE V
Change in item distribitions over time (n = 231)

	Set "N" items			Set "P" items		
	Code 1%	Code 2%	Code 34%	Code 1%	Code 2%	Code 34%
Wave 1	55	32	13	21	68	11
Wave 2	60	33	7	19	74	7
Wave 3	62	30	8	18	74	8
Wave 4	65	29	6	17	77	6

TABLE VI
Scores over time

	Wave 1	Wave 2	Wave 3	Wave 4
(a) High Scores				
GHQ: % scoring 5+	26.9	15.1	17.7	12.6
CGHQ: % scoring 12+	26.9	21.3	22.5	17.8
(b) Mean over Wave 1 mean				
GHQ	1.00	.59	.68	.50
CGHQ	1.00	.83	.82	.73

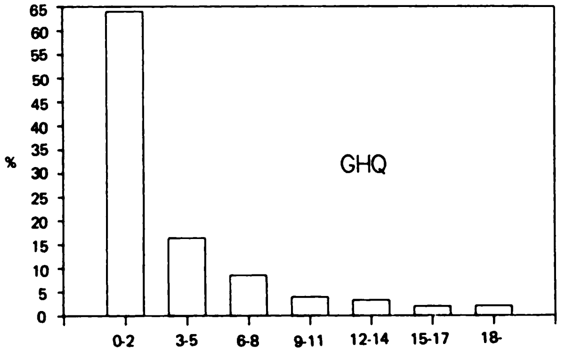

FIG 1a.—Frequency distribution for conventional GHQ score

tion of the GHQ improves detection of the more specific symptoms.

Relationships of the sub-scores to caseness were also investigated. The sensitivity and specificity of the main sub-scores were: P34, 77%/68%: N34, 72%/70%; and N234, 82%/89%. N1 and P2 had the expected negative relationships to caseness, but P1 was virtually independent of caseness. When N2 and N34 are entered into the regression as two separate scores, N2 only receives about two-thirds of the weight given to N34.

Longitudinal analysis and further validation

Henderson *et al* (1981) noted that these data show quite a pronounced re-test effect: the apparent level of neurotic illness falls from one interview to the next. In a general population sample, this is unexpected, and the reasons for it are not clear. Similar effects may be observed in other studies, and several hypotheses can be advanced to explain them. The re-test effect can be seen clearly in Table V, which shows the item marginal distribution at each interview, averaged over the 15 items in each set. In set C, the use of code 1 increases steadily from wave to wave, the use of codes 3 and 4 falls dramatically, and the use of code 2 falls quite slightly. In set P, code 2 increases and the other codes fall.

The effect of these shifts on the total scores is shown in Table VI. The upper half of the Table shows the percentage above fixed cutting-points that initially define the top 27%. The lower half of the Table shows the mean scores at each

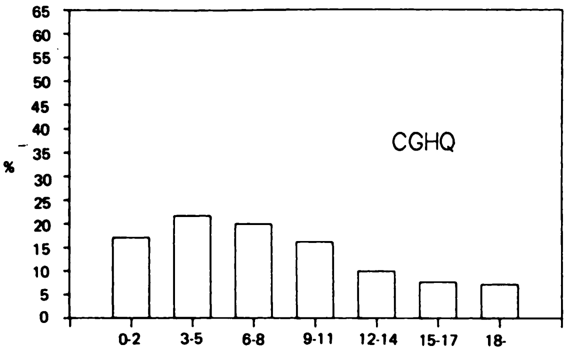

FIG 1b.—Frequency distribution for revised GHQ score

https://doi.org/10.1192/bjp.146.1.55 Published online by Cambridge University Press

wave, divided by the mean scores at wave 1. The CGHQ shows a quite definite re-test effect, but it is much weaker than the pronounced effect shown by GHQ.

Given the overall changes in use of answer categories exhibited in Table V, it seems worth enquiring whether there are specific patterns of movement from one answer category to another. We investigated this in two ways. Firstly, we examined the changes between waves 1 and 2, since this is where the greatest gross change took place. For each item, we cross-tabulated the wave 1 codes against the wave 2 codes, and pooled these cross-tabulations to give a single summary table for the N set and a similar one for the P set. As might be expected, these tables showed a degree of stability in the use of each category, indicated by concentration in the diagonal cells. This can be indexed by the ratio of the observed to the expected value in each diagonal cell, O/E, using the usual chi-squared calculation for a two-way contingency table. These ratios were: N1, 1.4; N2 1.6; N34 3.4; P1 2.0; P2 1.1; and P34 2.9; they show that the use of code 2 ("as usual") in the P set is less stable over four months than the use of the other codes. Standardised residuals (O–E, divided by the square root of E) tell a similar story. For category 2 in the P set, the standardised residual is 4.4; the other five range from 10 to 13.

We next tested for the presence of specific patterns of change from one code to another. Given the degree of stability in the use of the same code on both occasions, this is best done by fitting Goodman's (1965, 1979) model of quasi-independence (Bishop *et al,* 1975, equation 5.3-1; Jones & Pittelkow, 1983). This model supplies expected values for the off-diagonal cells, given the marginal totals, and assuming the diagonals are fixed at their observed values. The quasi-independence model gave an excellent fit in both the N and P sets, suggesting that there are no *specific* patterns of movement from one response category to another, *beyond* what would be expected from the marginal distributions, after allowing for stability in the use of each category.

Our second approach to the analysis of stability and change used rank correlations between sub-scores at one time-point and the next. The previous analysis focussed on stability in the use of specific item responses. The present one looks at stability in the use of a particular type of response category. Over the four waves, there are three four-month stability correlations for each score. The *averages* of these are: N2 0.61; N34 0.40; N234 0.68; P1 0.42; P2 0.50; and P34 0.38.

We turn now to the relationship of the GHQ and CGHQ to variables that are only available in this sub-set of the data. The DSSI state anxiety/state depression index correlated 0.48 with the GHQ and 0.64 with the CGHQ (average of Spearman correlations from waves 3 and 4). The life event score correlated 0.35 with the GHQ and 0.29 with CGHQ (average of correlations from all four waves). For the personality trait of neuroticism (EPI N score), we calculated the correlation with the total GHQ scores, summed across the four waves, since we are now examining a persistent dispositional characteristic. These correlations were: GHQ 0.47 and CGHQ 0.65. The correlations of neuroticism with the summed sub-scores are: N2 0.57; N34 0.49; N234 0.65; P1 0.05; P2−0.28; and P34 0.38. Evidently, trait neuroticism is more strongly related to the C items than to the P items. The correlations with N, wave by wave, are also of interest; they are: GHQ 0.39, 0.40, 0.28, and 0.24; CGHQ 0.55, 0.57, 0.52, 0.57. The GHQ correlations fall off in the later waves; those for the CGHQ are consistently higher, and remain steady.

Discussion

The GHQ is quite possibly the best instrument of its kind that we have, and since its continued widespread use seems assured, attention to details of scoring is appropriate. We have proposed a revised scoring for the GHQ, giving greater weight to indications of chronicity, and believe this scoring will give a more accurate index of the person's 'present state'. In our data base, the revised scoring has usefully greater validity, in the sense that it correlates more highly than the conventional scoring with other questionnaire measures of neurotic symptomatology, and gives better prediction of 'caseness', as defined by a standardised psychiatric interview. It correlates more highly with trait neuroticism, but somewhat less strongly with life-events; this slightly lower correlation is realistic, if the revised scoring does indeed represent both chronic and acute illness. It constitutes evidence of construct validity.

The sensitivity and specificity values given above are lower than most previously reported for the GHQ. While the reasons for this are not certain, it is likely to be due mainly to the time interval (mean 4.7 days) between the GHQ and the PSE.

Whether the new scoring will be universally appropriate depends on the reasons for administering the GHQ. An instrument of this kind may be used for screening in clinical settings, or for preventive health care, and it also has a variety of research uses. The latter include case-finding, epidemiological surveys (where the score may be used as a measure of pathology in its own right), and evaluation of intervention trials. In each of these contexts, one has to ask whether it is more relevant to obtain a measure that covers the range of illness conditions, or a measure of acute dysphoria, which may be confounded with ephemeral, and possibly appropriate reactions to transient stress. We believe a measure that gives a more accurate reflection of chronic, as well as acute conditions will often be the better choice.

The revised scoring has a number of other advantages. It has much more satisfactory statistical properties, with good discrimination over the whole range, and a more nearly normal distribution (Figure 1); standard statistical procedures can be applied to it with fewer qualms. It also has a greater stability in repeated use over time. The correlation with trait neuroticism remains constant, and the drop in mean score from one administration to the next is much less than with the conventional scoring. It thus provides a more robust tool for evaluation research, though since there is still a clear re-test effect, it should not be used in '*before and*

https://doi.org/10.1192/bjp.146.1.55 Published online by Cambridge University Press

after' designs without a control group. There are, of course, other reasons to question whether the no-control-group design has a valid role in evaluation studies (Campbell & Stanley, 1973). In intervention trials, it will provide a more conservative, but probably more realistic result.

The present investigation may throw some light on the reasons for re-test effects. This is a complex issue, which we shall discuss in greater detail elsewhere. However, the present findings are compatible with a view that the second or later response to a questionnaire is more considered, paying closer attention to the exact meaning of the questions and response categories, and perhaps to the time-frame.

We have demonstrated a specific pattern in the use of 'as usual' responses that we believe to be a valid reflection of chronic illness. An alternative hypothesis is that this pattern is a result of response style. A person given to careful, cautious answers might choose to respond 'no more than usual', where others would reply 'not at all'. This explanation does not fit well with our detailed findings, First, the 'as usual' response is much less common for the N items than for the P items. Secondly, the N2 responses correlate positively with other measures of neurotic illness, while the P2 responses correlate negatively with the same measures. There may be an element of response style effect in the use of the N2 response, but we believe the chronicity interpretation is more consonant with our findings.

Our investigation has uncovered a series of related anomalies in the use of P1 responses—answering 'better than usual' to positive items. The use of this response is uncorrelated with other measures of illness and with trait neuroticism. Yet it has been regarded as the most healthy response, so that significant negative correlations would have been expected. This was an unanticipated finding, for which we have no very plausible explanation, and consequently should be treated with some reserve until it is confirmed by others. The finding casts doubt on the appropriateness of the Likert scoring for the GHQ.

We hope *all* our present findings will be tested by others; fortunately, this is cheap and easy, since many workers have substantial GHQ data-bases. Until refutation or confirmation of our results becomes available, the revised scoring should be adopted for most future research, and may also be used to re-interpret previous studies. The revised scoring should be equally applicable to the 28-item GHQ (Goldberg, 1978), and should have a greater impact than on the 30-item version studied here, since three-quarters of the items in the 28-item version are negative, and subject to re-scoring. Our findings provide pointers to a more refined scoring system, perhaps along the following lines:

N1	0	P1	1
N2	2	P2	0
N34	3	P34	3

This takes account of the ambiguous status of the P1 responses and the fact that (in our data) the N2 responses are a little less predictive of 'caseness' than the N34 responses. However, this could only be recommended after testing in a very much larger database.

This investigation may also point to a wider conclusion for the construction of tests and instruments. Test contructors almost always rely heavily on their own and their colleagues' perceptions of common linguistic usage; detailed systematic investigation of the meaning and interpretation of questions and answer categories is the exception. Yet methods for testing the comprehension and understanding of survey questions have been developed (Belson, 1962, 1981; Gordon, 1963), and the application of these methods has often uncovered unanticipated ambiguity. This is arduous, painstaking work, but may be justified when a test or instrument is intended for very extensive use.

Acknowledgements

This investigation was prompted by comments from E. Jane Ellis, while entering GHQ data at a computer terminal.

References

Belson, W. A. (1962) *Studies in Readership*. London: Business Publications.

—— (1981) *The Design and Understanding of Survey Questions*. England: Gower.

Bishop, Y. M. M., Fienberg, S. E. & Holland, P. W. (1975) *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, Mass: MIT Press.

Campbell, D. T. & Stanley, J. C. (1963) *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rank McNally College.

Chan, D. W. & Chan, T. S. C. (1983) Reliability, validity and the structure of the General Health Questionnaire in a Chinese context. *Psychological Medicine*, **13**, 363–71.

Duncan-Jones, P. (1981) The natural history of neurosis: Probability models. In: *'What is a Case? The Problem of Definition in Psychiatric Community Surveys'*. (eds. J. K. Wing, P. Bebbington & L. N. Robins) London: Grant McIntyre.

Efron, B. (1978) Regression and ANOVA with zero-one data: Measures of residual variation. *Journal of the American Statistical Association*, **73**, 113–21.

Eysenck, H. J. & Eysenck, S. B. G. (1964) *Manual of the Eysenck Personality Inventory*. London: London University Press.

Finlay-Jones, R. A. & Murphy, E. (1979) Severity of psychiatric disorder and the 30-item General Health Questionnaire. *British Journal of Psychiatry*, **134**, 609–16.

Foulds, G. A. (1976) *The Hierarchical Nature of Personal Illness*. London: Academic Press.

Goldberg, D. (1972) *The Detection of Psychiatric Illness by Questionnaire. Maudsley Monograph, No. 21*. London: Oxford University Press.

—— (1978) *Manual of the General Health Questionnaire*. Windsor: NFER-Nelson.

GOODMAN, L. A. (1965) On the statistical analysis of mobility tables. *American Journal of Sociology*, **70**, 564–85.

—— (1979) Multiplicative models for the analysis of occupational mobility tables and other kinds of crossclassification tables. *American Journal of Sociology*, **84**, 804–19.

GORDON, W. D. (1963) Double interview. In: *'New Developments in Research.'* London: Market Research Society with the Oakwood Press.

HARDING, T. W. (1976) Validating a method of psychiatric case identification in Jamaica. *Bulletin of the World Health Organisation*, **54**, 225–31.

HENDERSON, S., BYRNE, D. G. & DUNCAN-JONES, P. (1981) *Neurosis and the Social Enviroment*. Sydney: Academic Press.

JONES, F. L. & PITTELKOW, Y. E. (1983) Analysis of occupational mobility tables using GLIM. *GLIM Newsletter*, **7**, 34–6.

LILIENFELD, A. M. (1976) *Foundations of Epidemiology*. London: Oxford University Press.

MUNOZ, P. E., VAZQUEZ, J. L., PASTRANA, E., RODRIGUEZ, F. & ONECA, C. (1978) Study of the validity of Goldberg's 60-item GHQ in its Spanish version. *Social Psychiatry*, **13**, 99–104.

WING, J. K. (1976) A technique for studying psychiatric morbidity in inpatient and outpatient series and in general population samples. *Psychological Medicine*, **6**, 665–72.

—— COOPER, J. & SARTORIUS, N. (1974) *The Measurement and Classification of Psychiatric Symptoms*. Cambridge: Cambridge University Press.

YERUSHALMY, J. (1947) Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques. *Public Health Reports*, **62**, 1432–99.

ZUNG, W. W. K. (1965) A self-rating depression scale. *Archives of General Psychiatry*, **12**, 63–70.

M. E. Goodchild, BA, *Research Assistant*

*Paul Duncan-Jones, MA, *Senior Research Fellow*

National Health and Medical Research Council Social Psychiatry Research Unit, Australian National University, P.O. Box 4, Canberra City, ACT 2601, Australia

*Correspondence.

(Received 28 February; revised 1 May 1984)

https://doi.org/10.1192/bjp.146.1.55 Published online by Cambridge University Press