THE GENERAL HEALTH QUESTIONNAIRE: RELIABILITY AND VALIDITY FOR AUSTRALIAN YOUTH

Helen R. Winefield, Robert D. Goldney, Anthony H. Winefield and Marika Tiggemann

General Health Questionnaire (GHQ) results are given for a large (N = 1013) sample of South Australian young people (average age 19.6 years), to compare the usefulness of the 12-, 28-, and 30-item forms of the GHQ. Internal reliabilities are generally adequate and the Likert scoring method produces significant correlations with psychological measures such as self-esteem. The case-prevalence rate using the binary scoring method was comparable with other studies, but misclassification rates were unacceptably high when DSM-III Axis I diagnosis was used as the criterion for the presence of any psychiatric disorder.

Australian and New Zealand Journal of Psychiatry 1989; 23: 53-58

The General Health Questionnaire (GHQ) [1-3] has been widely used as a psychiatric screening instrument in a variety of community populations. These include general practice patients [4-7], working adults [8, 9], and sufferers of chronic illnesses and disabilities [10-12]. Although the reliability, validity, and acceptability to subjects of the GHQ have been convincingly demonstrated, questions remain concerning the utility for younger community subjects of the several forms of different length which are available. For example Banks *et al.* [8] found the briefest (12-item) form provided a useful index of psychiatric illness for the study of employment and related issues, and advocated its use there on the grounds of typical time pressures. On the other hand Burvill and Knuiman [5] advised using the longest (60-item) form for community samples on several grounds, including its low misclassification rates.

To compound the uncertainties of researchers contemplating use of the GHQ in community samples, there are two well-established scoring methods. One method is to reduce the 4 possible answers to each item to a binary 0/1 score, then to report the prevalence rate for psychiatric disorder based on the percentage of subjects whose resultant total score reaches some critical threshold or cut-off level. For example the 4/5 cut-off usual for the 28-item version means that subjects with binary-method scores of 5 or higher are regarded as psychiatrically disturbed [3]. Bridges and Goldberg [10] found however that in a population of neurological in-patients, a GHQ-28 cut-off point of 11/12 was more appropriate. They reached this conclusion from a Receiver Operating Characteristic analysis to see which cut-off point maximized both sensitivity (the proportion of true positives: independently diagnosed cases which the test correctly identifies) and specificity (the proportion of true negatives: non-cases which the test correctly identifies). The alternative

Department of Psychiatry, University of Adelaide
Helen R. Winefield, PhD
Dibden Research Unit, Glenside Hospital
Robert D. Goldney, MD
Dept. of Psychology, University of Adelaide
Anthony H. Winefield, PhD
School of Social Sciences, Flinders University
Marika Tiggemann, PhD
Correspond with Dr. H.R. Winefield, Dept. of Psychiatry, University of Adelaide, GPO Box 498, Adelaide, South Australia 5001.

Downloaded from anp.sagepub.com at PENNSYLVANIA STATE UNIV on September 16, 2016

scoring method is to employ a Likert-style 0, 1, 2, or 3-point allocation for each item. This method results in totals with a wider variability, where the mean and standard deviation allow study of differences amongst low-prevalence groups with greater precision than does the traditional binary scoring method [13].

The sensitivity and specificity of the GHQ have been established through comparisons with diagnoses of psychiatric disorder which derive from standardized interview-based techniques such as the Present State Examination [11-13], the International Classification of Diseases [6], and Goldberg's Clinical Interview Schedule [7, 10]. However, there do not yet appear to be reports of its validation with the DSM-III diagnostic classification system [14], which is the nosological system most frequently employed in Australian clinical practice.

The need for a reliable, valid, and yet brief self-report measure of psychological adjustment, which is suited to young adults, becomes pressing in investigations of the transition from secondary school into the workforce. In the course of a major longitudinal study of over 3000 South Australians, beginning when their average age was 15.6 years and continuing for almost a decade, we have had the opportunity to collect data on the psychometric properties of the GHQ when used with this population.

In this paper we shall first discuss the properties and relative psychometric merits of the 12-, 28-, and 30-item forms of the scale. As the psychological and physical health correlates of unsatisfactory work have been shown to be as unfavourable as those of unemployment in this population [15, 16], norms are needed for dissatisfied employed and unemployed in addition to norms for satisfied employed and full-time tertiary education students. The effects on GHQ score of sex, socio-economic status, and cultural background also require assessment, although most authors have found the measure relatively insensitive to these possible influences [5, 7]. Concurrent validity is investigated here through the correlations between GHQ and other measures of psychological state, based on both self-report and a psychiatric interview following DSM-III guidelines.

Method

Subjects

Subjects were 1013 young people, with average age in 1984 of 19.6 years (SD = 1.04). In 1980 they had been students at 12 randomly-chosen metropolitan secondary schools, and they had taken part thereafter in four annual postal follow-ups which obtained an average 76% return-rate. Retrospective analyses of those who discontinued participation have shown that they did not differ systematically from continuing participants on any of the variables measured. In the sample reported here, 48.2% of subjects were male, the socio-economic status of fathers' occupations was categorized as high for 36.7%, medium for 38.6%, and low for 24.7% [17], and a language other than English was regularly spoken at home for 17.9% (non-Anglo immigrant background). Work status was divided into four categories: full-time student (23.7%), unemployed (7.7%), satisfied employed (61.1%), and dissatisfied employed (7.6%), the latter two groups being distinguished according to Warr *et al.* Job Satisfaction Scale [15, 18]. One year later (in 1985) 729 of the same subjects completed questionnaires, and in 1986, 623 did so.

A sub-sample of the 1984 respondents was selected to undergo more intensive study by an interview with one of five (four male, one female) experienced clinical psychiatrists. The members of this sub-sample were (a) GHQ-positive cases (using the 4/5 binary score criterion on the 28-item form of the GHQ), and (b) matched controls. Altogether there were 291 GHQ-positive cases, of whom 118 were interviewed on the basis of availability. Those interviewed did not differ from the other GHQ-positive subjects in sex distribution, but did include more students and fewer satisfied employed (Chi-square = 23.27, df = 3, $p < 0.001$). For each GHQ-positive subject interviewed, a control was chosen by selecting the next GHQ-negative (non-case) subject of the same sex from the same original school-class list. Thus sub-sample cases and non-case controls were matched for sex, age and school attended (an indicator of socio-economic status), the simplest way of avoiding obvious systematic differences between the groups.

Measures

In 1984 subjects responded to 44 GHQ items which incorporated the 12-, 28-, and 30-item forms [1-3], Rosenberg's Self-Esteem and Depressive Affect Scales [19], and the Nowicki-Strickland Internal-External Locus of Control Scale [20].

For the interviewed sub-sample, psychiatrists com-

Downloaded from anp.sagepub.com at PENNSYLVANIA STATE UNIV on September 16, 2016

Table 1. Some psychometric characteristics of 3 versions of the GHQ in Australian youth

	12-item	28-item	30-item
Males (max. n=485)			
mean	10.74	18.12	23.66
(S.D.)	(4.76)	(10.42)	(11.02)
alpha	0.84	0.92	0.92
1 yr. retest	*0.43	*0.44	*0.43
2 yr. retest	*0.41	*0.50	*0.45
Females (max. n=521)			
mean	10.65	18.38	22.68
(S.D.)	(4.94)	(10.19)	(10.86)
alpha	0.84	0.91	0.92
1yr. retest	*0.34	*0.38	*0.35
2 yr. retest	*0.22	*0.32	*0.25

*Pearson r correlation, p<0.001

pleted the Global Assessment Scale [21] and judged the evidence for an Axis I diagnosis of psychiatric disorder using the DSM-III as definite, probable, equivocal, possible, or none. Inter-rater reliability training using videotaped interviews, resulted in an average agreement coefficient of .75 between the five psychiatrists over the Axis I items, using Fleiss's modification of kappa for more than two raters [22]. This was considered to be satisfactory for the present study.

Procedures

Self-report measures with less than a 90% completion rate of component items were treated as missing data, leading to slightly different subject numbers in different sections. Norms, internal reliability (alpha) coefficients, test-retest correlations over one and two years, and the effects of background variables are given for all the 1984 respondents. In addition, concurrent validity of the GHQ is studied by examining its relationship both with self-report psychological measures, and with judgements by psychiatrists blind to the GHQ status of each subject. Psychiatric interviews were carried out within 3 months of subjects' completing the self-report measures including the GHQ. Both Likert and binary scoring methods have been used, the former to provide an interval-scale score, and the latter to distinguish cases of psychiatric disorder, using various cut-off points.

Results

Means and Reliability

Means, SDs, alpha coefficients to show internal consistencies, and test-retest correlations over one and two years, are shown in Table 1 for each of the 12-item, 28-item, and 30-item forms, separately for males and females, and all using Likert scoring. Alpha reliability coefficients showed a similar pattern in the later years: for males in 1985 and 1986 alpha was .83 and .78 for the 12-item form, .90 and .88 for the 28-item form, and .91 and .89 for the 30-item form; for females the corresponding figures were .86 and .84, .92 and .91, and .93 and .91.

Scores for the 12-item form correlated .87 with the 28-item and .96 with the 30-item form, while the latter two forms correlated with each other at .91 (Pearson product-moment correlations, $p<0.001$).

Internal reliability of the 12-item form is somewhat lower than for the two longer forms, although still at an acceptable level. The stability of scores for the 28-item form is slightly higher than for the other forms, especially over two years. All forms are consistently higher in stability for males than for females.

Effects of background variables

There were no statistically significant differences in scores for males and females, and accordingly data will be presented for the combined sexes for the remainder of the paper, unless otherwise noted.

Scores did not differ by socio-economic class, for males or females. However for females but not for males, a non-English speaking (immigrant) home was associated with higher scores than English-only (12-item means 11.8 and 10.4, 28-item means 21.1 and 17.8, 30-item means 25.8 and 22.0; all significant at $p< 0.05$ by t-test).

Differences between occupational groups are shown in Table 2. The satisfied employed scored significantly lower than the other three occupational groups (Fs (3, 919) = 19.76, 18.45, and 20.82 for the 12-item, 28-item, and 30-item forms respectively, for all of which $p< 0.001$).

Downloaded from anp.sagepub.com at PENNSYLVANIA STATE UNIV on September 16, 2016

Table 2. Mean GHQ scores by occupational group, for three forms of the test

	Satisfied employed (n = 566)	Dissatisfied employed (n = 70)	Unemployed (n = 71)	Students (n = 220)
12-item	9.74	12.19	12.28	12.04
28-item	16.32	22.41	22.08	20.18
30-item	20.91	27.16	27.23	26.00

Correlations with psychological measures

GHQ scores correlated significantly ($p< 0.001$) with other self-report psychological measures as follows: with self-esteem r = -.44 (12-item), -.47 (28-item), and -.46 (30-item); with (external) locus of control .24 (12-item), .31 (28-item), and .28 (30-item); with depressive affect .54 (12-item), .55 (28-item), and .57 (30-item).

GHQ-positive cases

The GHQ-positive case rate using the traditional 4/5 criterion of the 28-item binary form was 28.7% at the first testing, 20.3% one year later, and 27.6% two years later. Of those who were cases at the first testing, 36.9% were cases in the following year, and 44.9% were still cases two years later.

At the initial assessment, there was no difference in case rates between males and females, subjects of different socio-economic status, or with immigrant background or not. Again, however, there was an association between occupational group and classification as a case, with fewer of the satisfied employed (21.1%) being cases than the dissatisfied employed (39.7%), the unemployed (39.7%), or the students (39.1%) (Chi-square=37.14, df=3, $p<0.001$)

Interviewed sample of cases and matched controls

Interviewed GHQ-positive cases had been matched with controls in sex, age, and socio-economic background. However for the females significantly more cases than controls had a non-English speaking immigrant background (33.9% vs. 12.5%; Chi-square = 6.29, df = 1, $p< 0.02$).

Interviewer adjustment ratings using the Global Assessment Scale were lower for cases than for controls, means 76.4 vs. 83.7 ($t(223) = 4.63$, $p< 0.001$).

Psychiatrists judged 31.4% of the GHQ-positive cases and 10.2% of the matched controls (GHQ-negative non-cases) to show some evidence of DSM-III Axis I disorder. There was a significant association between being a GHQ-positive case and being diagnosed as showing any evidence of Axis I disorder (Chi-square = 15.42, df = 1, $p< 0.001$), for males and females combined. Dividing the five categories of probability of psychiatric disorder differently, for females 29.5% of GHQ cases and 7.1% of controls were classified as having "probable" or "definite" evidence of Axis I disorder (Chi-square = 15.07, df = 1, $p< 0.001$), but for males only 19.3% of GHQ cases, and 7.7% of controls, were classified as having "probable" or "definite" evidence of Axis I disorder (Chi-square = 3.16, df = 1, $p> 0.05$).

Taking the psychiatric interview as the diagnostic criterion, the GHQ misclassification rate is the ratio of the sum of false negative cases (i.e. GHQ-negative, interview-positive) plus false positives (i.e. GHQ-positive, interview-negative), to the total number of subjects. With the 4/5 cut-off point, the misclassification rate was 45.0% for males and 36.8% for females. Sensitivity (the ratio of GHQ-and-interview-positives to all interview-positive cases) was .70 for males and .82 for females. Specificity (the ratio of GHQ-and-interview-negatives to all interview-negatives) was however lower, at .52 for males and .57 for females.

As the binary-scored GHQ-28 had low specificity using the 4/5 criterion of psychiatric disorder, we recalculated its validity against the DSM-III using a more stringent cut-off at 7/8. Cases defined in this way differed from non-cases on the Global Assessment Scale by a similar margin to when cases and non-cases were defined using the 4/5 cut-off, means 75.0 and

Downloaded from anp.sagepub.com at PENNSYLVANIA STATE UNIV on September 16, 2016

81.9 respectively, $t(223) = 3.88$, $p < 0.001$. Application of the 7/8 criterion meant that 36.9% of all GHQ cases, and 14.9% of non-cases, were diagnosed as showing any evidence of psychiatric disorder using the DSM-III interview. As expected, GHQ specificity increased, to .72 for males and .82 for females; at the same time sensitivity inevitably fell, to .45 for males and .54 for females. Overall misclassification rates fell to 33.0% for males and 24.8% for females.

Discussion

Mean GHQ Likert scores in the present community sample of late adolescents are comparable with but somewhat higher than those from Banks' study of 200 male and female 17 year-olds [13], half of whom were unemployed. Banks *et al.* [8] reported means of 8.67 for employed and 14.06 for unemployed 16 year-olds using the 12-item form. Donovan *et al.* [23] followed up 16-year old school leavers and found GHQ-12 mean scores of 6.45 for employed and 11.00 for unemployed males. Higher mean scores (35 on GHQ-30 and 15 on GHQ-12) have been reported for middle-aged unemployed samples [9, 24].

Our results confirm those of other investigators concerning the impressive internal reliability of the GHQ [8]. The implication of this is that all the items measure something similar, and any subscales are highly intercorrelated and probably add little extra information. Relatively lower test-retest coefficients over one and two years are not surprising given the challenges and developmental tasks being faced at this period of early adulthood, and thus do not reflect on the test's reliability. However it is of interest that stability of scores over time appears to be lower for female than for male subjects, leading to speculations of greater psychological fluidity in females of this age group compared with males. This is consistent with the greater increase in self-esteem of girls who obtained jobs compared with boys [25, 26].

GHQ correlations with self-reported self-esteem, locus of control, and depressive affect are moderate and show little difference between the three forms.

Using the standard 4/5 criterion of the GHQ-28 as the cut-off point, the case/prevalence rate in our sample (40% for the unemployed and 23% for the combined employed groups) is comparable to but somewhat lower than the results of McPherson and Hall [27] who, using the GHQ-12, found a 48% case rate with unemployed Sydney 17 year-olds and 28% for 19 year-old apprentices. Boardman [4] found 43% cases with 16-19 year-olds and 49% with 19-29 year-olds, with the GHQ-28 in a UK general practice sample; Finlay-Jones and Eckhardt [28] reported a rather higher case rate of 56% for unemployed 16-24 year-olds in Canberra, using the GHQ-30. The reason for the discrepancy between our findings and those of Finlay-Jones and Eckhardt is not immediately clear. Our results concur with other researchers who have studied both employed and unemployed youth, in finding about twice the prevalence of cases amongst the unemployed. An important further point to note is the value of investigating job satisfaction as a mediating variable, as our dissatisfied employed subjects showed an equivalent case rate to the unemployed.

A strength of the GHQ is its apparent resistance to the effects of sex, socio-economic status, and cultural background except for the higher scores of immigrant-background females. In this sample, immigrant-background females were particularly likely to have entered into tertiary education [29]. The higher scores and higher case rate of students compared to other occupational groups needs follow-up investigation, especially in the light of Boardman's [4] finding with GHQ-28 that students had lower case rates than the employed, and half that of unemployed, whereas in our sample unemployed and students were equivalent. A useful future exploration would be to follow up the psychological adjustment of our student group after they have joined the workforce - perhaps before that they are showing the same failure to increase in well-being that non-workers did at earlier ages [26].

It does not seem that any one form of the GHQ is clearly "the best" for young Australians. Patterns are similar with all three forms; where the 12-item gains in brevity it loses slightly in having lower alpha, lower 2-year retest coefficient (especially for females), and slightly lower correlations with psychological measures. Researchers will need to choose the form which best suits their purposes and sample. The Likert scoring method is valuable in increasing the range of variability in scores and making them more suitable for statistical analyses requiring interval-scale data.

The ultimate criterion for GHQ validity has always been regarded as its concordance with the results of a psychiatric interview. In the present study the GHQ misclassified a high proportion of interviewed subjects, especially in the direction of low specificity (too many false positives) when using the 4/5 cut-off point. This effect was particularly marked for male subjects,

Downloaded from anp.sagepub.com at PENNSYLVANIA STATE UNIV on September 16, 2016

a finding consistent with that of Tarnopolsky *et al.* [6]. Raising the cut-off point to 7/8 classified fewer subjects as GHQ "cases" and therefore improved specificity, but also decreased the detection rate for true cases (sensitivity). Because more of the subjects were non-cases, raising the cut-off point reduced the total misclassification rate; however individual users of the GHQ will have to decide which sort of error is more important to avoid: failing to detect true cases, or classifying too many non-disturbed people as cases. In our data cut-off points from 4/5 to 7/8 showed a direct trade-off between sensitivity and specificity.

The psychiatric interviews in this study were performed by private practitioner rather than research psychiatrists. Although inter-rater reliability was acceptable, the question arises of whether interviewers' perceptions were biassed by their clinical experiences towards overlooking less severe psychopathology.

It is also possible that the American-derived DSM-III diagnoses are more austere in their definition of illness than are the British-based diagnostic instruments with which the GHQ has been validated in the past. At the very least it would appear that the GHQ identifies more cases than do Australian clinicians using the DSM-III, and as the DSM-III is widely used in Australia, this observation is worthy of further research.

References

1. Goldberg D. The detection of psychiatric illness by questionnaire. Maudsley Monograph No. 21. Oxford University Press, 1972.
2. Goldberg D. Manual of the General Health Questionnaire. Slough: National Foundation for Education Research, 1978.
3. Goldberg DP, Hillier VF. A scaled version of the General Health Questionnaire. Psychological Medicine 1979; 9: 139-145.
4. Boardman AP. The General Health Questionnaire and the detection of emotional disorder by General Practitioners: A replicated study. British Journal of Psychiatry 1987; 151: 373-381.
5. Burvill PW, Knuiman MW. Which version of the General Health Questionnaire should be used in community studies? Australian and New Zealand Journal of Psychiatry 1983; 17: 237-242.
6. Tarnopolsky A, Hand DJ, McLean EK, Roberts H, Wiggins RD. Validity and uses of a screening questionnaire (GHQ) in the community. British Journal of Psychiatry 1979; 134: 508-515.
7. Tennant C. The General Health Questionnaire: A valid index of psychological impairment in Australian populations. Medical Journal of Australia 1977; 2: 392-394.
8. Banks MH, Clegg CW, Jackson PR, Kemp NJ, Stafford EM, Wall TD. The use of the General Health Questionnaire as an indicator of mental health in occupational studies. Journal of Occupational Psychology 1980; 53: 187-194.
9. Jackson PR, Warr PB. Unemployment and psychological ill-health: The moderating role of duration and age. Psychological Medicine 1984; 14: 605-614.
10. Bridges KW, Goldberg D. The validation of the GHQ-28 and the use of the MMSE in neurological in-patients. British Journal of Psychiatry 1986; 148: 548-553.
11. Lindesay J. Validity of the General Health Questionnaire (GHQ) in detecting psychiatric disturbance in amputees with phantom limb pain. Journal of Psychosomatic Research 1986; 30: 277-281.
12. Rabins PV, Brooks BR. Emotional disturbance in multiple sclerosis patients: Validity of the General Health Questionnaire (GHQ). Psychological Medicine 1981; 11: 425-427.
13. Banks MH. Validation of the General Health Questionnaire in a young community sample. Psychological Medicine 1983; 13: 349-353.
14. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, Third Edition. Washington, D.C.: APA, 1980.
15. Winefield AH, Tiggemann M, Goldney RD. Psychological concomitants of satisfactory employment and unemployment in young people. Social Psychiatry and Psychiatric Epidemiology 1988; 23: 149-157.
16. Winefield HR, Winefield AH, Tiggemann M, Smith S. Unemployment, drug use, and health in late adolescence. Psychotherapy and Psychosomatics 1987; 47: 204-210.
17. Broom L, Jones F. Career mobility in three societies: Australia, Italy, and the United States. American Sociological Review 1969; 34: 650-658.
18. Warr PB, Cook JD, Wall TD. Scales for the measurement of some work attitudes and aspects of psychological well-being. Journal of Occupational Psychology 1979; 52: 129-148.
19. Rosenberg M. Society and the Adolescent Self-Image. Princeton: Princeton University Press, 1965.
20. Nowicki S, Duke MP. A locus of control scale for noncollege as well as college adults. Journal of Personality Assessment 1974; 38: 136-137.
21. Endicott J, Spitzer RL, Fleiss JL, Cohen J. The Global Assessment Scale: A procedure for measuring overall severity of psychiatric disturbance. Archives of General Psychiatry 1976; 33: 766-771.
22. Fleiss JL. Measuring nominal scale agreement among many raters. Psychological Bulletin 1971; 76: 378-382.
23. Donovan A, Oddy M, Pardoe R, Ades A. Employment status and psychological well-being: A longitudinal study of 16-year old school leavers. Journal of Child Psychology and Psychiatry 1986; 27: 65-76.
24. Payne R, Warr P, Hartley J. Social class and psychological ill-health during unemployment. Sociology of Health and Illness 1984; 6: 152-174.
25. Gurney RM. Does unemployment affect the self-esteem of school-leavers? Australian Journal of Psychology 1980; 32: 175-182.
26. Tiggemann M, Winefield AH. The effects of unemployment on the mood, self-esteem, locus of control and depressive affect of school leavers. Journal of Occupational Psychology 1984; 57: 33-42.
27. McPherson A, Hall W. Psychiatric impairment, physical health and work values among unemployed and apprenticed young men. Australian and New Zealand Journal of Psychiatry 1983; 17: 335-340.
28. Finlay-Jones RA, Eckhardt B. A social and psychiatric survey of unemployment among young people. Australian and New Zealand Journal of Psychiatry 1984; 18: 135-143.
29. Winefield HR, Winefield AH, Tiggemann M, Goldney, RD. Psychological and demographic predictors of entry to tertiary education in young Australian females and males. British Journal of Developmental Psychology 1988; 6: 183-190.

Downloaded from anp.sagepub.com at PENNSYLVANIA STATE UNIV on September 16, 2016