Investigating Long-term Retest Effects in the GHQ-12

David J. Pevalin

Abstract

The aim of this analysis was to examine data from a general population sample for any retest effects in the 12-item General Health Questionnaire. A core panel was drawn from the British Household Panel Survey (n = 4749) of those who had completed the GHQ-12 seven times from 1991 to 1997. The panel results were compared with cross-sectional data from the Health Surveys for England for the same years. No evidence of retest effects was found. The age composition of the panel and the distribution of GHQ scores with age are discussed in light of these findings.

Non-Technical Summary

In this paper, we investigate whether or not the repeated application of the 12-item General Health Questionnaire (GHQ) results in any discernible retest effects. The GHQ has been administered to a core panel of the British Household Panel Survey (BHPS) once a year since 1991; seven times, up to and including wave 7. Other studies have found that the GHQ is liable to retest effects when administered multiple times over a short period, but it is uncertain if a longer time period between applications has similar results.

Results from the BHPS core panel are compared with data from the Health Surveys for England, a series of large cross-sectional studies conducted over the same years. Overall, the results indicate that no retest effects are present in the BHPS data and that the 12-item GHQ is a suitable measure of mental health for use in population based studies with relatively long time periods between applications.

Introduction

The General Health Questionnaire (GHQ) has been used as a screening instrument for minor psychiatric disturbance in numerous clinical studies as well as an indicator of psychiatric morbidity in large-scale, community-based surveys. The GHQ is usually self-administered and it is based on the respondent's assessment of their present state relative to their usual, or normal, state (Goldberg and Williams, 1988; Bowling, 1991).

Included in the studies and surveys that have employed the GHQ, a small number have had a longitudinal component. This has resulted in multiple completions of the GHQ by the same individual over the time of the study/survey. Using any instrument to measure change over time raises the possibility of the results being subject to retest, or panel conditioning, effects (Kalton and Citro, 1995). The most common being a 'social desirability' hypothesis where the respondent becomes familiar with the instrument over repeated completions and chooses the more 'acceptable' answers. This would result in lower GHQ scores. The competing hypothesis is that of increasing GHQ scores through an increased sensitivity to the questions.

Studies that have used the GHQ repeatedly report mixed results of any possible retest effects depending on the time between completions and overall time of the study. Relatively short follow-up periods (one year or less) are more common. Henderson et al. (1981) used the GHQ-30 in four waves over one year in a final sample of 231 from the general population of Canberra, Australia. They found significant falls in total GHQ scores and percentages of 5+ and 10+ threshold scores. They concluded that given the assumption of a stable state in the sample over the time of the study, the lower GHQ scores probably resulted from a ‘social desirability’ retest effect. Ormel et al. (1989) employed the GHQ-28 in a self-completion survey of a sample of new psychiatric outpatients in a Dutch city. The outpatients were studied over three waves in one year with a final sample of 175. They also found substantial retest effects in their sample. Kitamura et al. (1994) used the GHQ-30 on a sample of 120 pregnant Japanese women. The women completed the GHQ-30 four times over nine months starting in early pregnancy and ending at one month after the birth. They found that the GHQ-30 lost its validity against a diagnosis for the middle two applications but regained it for the last application one month after the birth.

Longer follow-up periods have been employed in two studies. Radovanovic et al. (1988) administered the full 60-item GHQ to 121 Yugoslav medical students three times with an interval of 2 years between tests. They found a marked decrease in mean GHQ-60 scores for males and females in each consecutive test. This was accompanied by a simultaneous decrease in sensitivity. Studies have also drawn on survey data. Graetz (1991) examined the GHQ-12 data from the Australian Longitudinal Survey over three years from 1985 to 1988 (4 waves). The data were collected yearly from an original sample of 8,998 young adults (16-25 years) and 6,151 respondents completed the GHQ in all four waves. While not primarily investigating retest effects, Graetz did find that the GHQ scores declined over time and hypothesised that retest effects might be one explanation.

In all of the studies reviewed above, the GHQ scores declined over time and multiple applications and any retest effects were thought to be operating through the ‘social desirability’ hypothesis. However, detecting retest effects in any instrument requires either a referent instrument or making assumptions about trends over time. With most large-scale surveys, there is no objective assessment of psychiatric impairment, such as a structured diagnostic interview, or even other instruments within the survey thus making any investigation of retest effects difficult. Also, over longer time periods a constant state in the population cannot be assumed as more macro effects, such as an economic recession, rising unemployment or job insecurity, may affect the mental health of the population as a whole.

In addition, the ageing of the panel members themselves has to be taken into account when using longer follow up times. The variation of the instrument with age is an important component in determining whether or not the instrument can be compared with first application and in which direction any effect of ageing should be. This is intertwined with the age of the sample as different ages could expect to be on different trajectories, especially if the variation with age is curvilinear.

This study investigates retest effects across seven applications (6 years) of the GHQ-12 in a general population sample by comparing the longitudinal results with those from a series of large-scale cross-sectional surveys that also used the GHQ-12 over the same time period.

Method

The data used in this study came from two sources: the British Household Panel Survey (BHPS) and the Health Surveys for England (HSE)[1]. The BHPS is an on-going annual panel survey of a

[1] The data used in these analyses come from (1) The British Household Panel Survey (BHPS). The BHPS is being conducted by the Institute for Social and

representative sample of more than 5,000 households from Great Britain. This results in approximately 10,000 individual interviews of adults aged 16 and over. Individuals of the initial sample of households at wave one - original sample members (OSM; n = 10,264) - continue to be followed even if they leave the original household. New individuals enter the panel if they move into a household containing an OSM, are born to an OSM, or an OSM moves into a household with one or more new people. The survey was first conducted in 1991 and this study used data from the first seven annual waves. Full details of the survey can be found in Buck (1990) and Taylor et al. (1998). The GHQ-12 formed part of the self-completion section in the BHPS.

From the BHPS, a balanced core panel was extracted. The OSMs in this panel had completed the GHQ-12 at all waves 1 through 7. Initially, the core panel included 5,513 but OSMs living in Scotland at wave 1 (n = 473), living in Wales at wave 1 (n = 257), or moved from England to Scotland or Wales (n = 64) were excluded to ensure comparability with the HSE data. This left a final panel of 4,749. The core panel was split into two panels by age at wave 1: 16-65 (n = 4,167) and 66+ (n = 582). The attrition rate obviously increases with age, mainly through death and illness, and splitting the panel in this way will reduce attrition bias in the younger panel. The younger panel contained 1,908 (45.7%) males and 2,259 (54.3%) females while the older panel had 234 (40.2%)

Economic Research (incorporating the ESRC Research Centre on Micro-social Change and the UK Longitudinal Studies Centre) at the University of Essex. (2) The Health Surveys for England (HSE) 1991-1995 and 1997. The Office for Population Censuses and Surveys carried out the HSE 1991-93. From 1994 it has been carried out by the Joint Survey Unit of the National Centre of Social Research and the Department of Epidemiology and Public Health at University College London. Both sets of data were provided by The Data Archive at the University of Essex. The analyses and interpretations presented in this paper are the sole responsibility of the author. This research was funded by the Health Education Authority under project no. UZ216.

males and 348 (59.8%) females. Non panel members living outside of England at any wave were excluded from the intra-BHPS comparisons.

The second data source were the HSEs 1991 to 1995 and 1997 (the GHQ-12 was not a part of the HSE 1996). The HSEs are a series of cross-sectional annual surveys that employ a representative sample for England. The 1991 and 1992 surveys had a sample of about 3,000 and 4,000 adults respectively. For 1993 to 1996 the adult sample was about 16,000 and in 1997 about 8,500. Full details of the surveys can be found in Prescott-Clarke and Primatesta, 1998 (HSE 1995-1997), Colhoun and Prescott-Clarke, 1996 (HSE 1994), Bennett et al., 1995 (HSE 1993), Breeze et al., 1994 (HSE 1992), and White et al., 1993 (HSE 1991). The GHQ-12 was also part of the self-completion section in the HSEs. Comparison groups derived from the HSE data increased with age to match the ageing of the panel. This, along with full completions of the GHQ-12, resulted in the samples reported in Table 1.

* * Table 1 about here * * *

The GHQ items from both data sources were coded in ordinal (0-1-2-3) format. This has a number of advantages over the bimodal (0-0-1-1) coding format for complex analyses. The ordinal coding resulted in an overall scale that ranged from 0 to 36.

The analyses were conducted in two parts. The first investigated whether or not GHQ scores differed among those BHPS respondents in the panel and out of the panel at wave 1 (1991) and wave 4 (1994). The second part was a comparison of the GHQ scores across both surveys broken down by sex and age group. Most other studies using the GHQ-12 have found that females have higher scores than males. Therefore all analyses were split by sex. While selection bias in the panel may result in consistently higher or lower GHQ scores any retest effects in the panel would show up by a divergence from the cross-sectional results.

Results

First, we tested to see if any there were any differences in GHQ scores at wave 1 for panel members and other OSMs in BHPS. In the younger panel no significant differences were found in mean GHQ score at wave 1 (males: panel mean 10.01, non-panel mean 10.26, *t* = 1.54; females: 11.07, 11.34, 1.51). However, in the older panel the non-panel members had significantly higher mean GHQ scores (males: panel mean 9.95, non-panel mean 10.94, *t* = 2.42; females: 10.50, 11.98, 4.35). Panel members were generally younger than non-panel members, especially in the older panel (males: panel mean 72.41, non-panel mean 74.54, *t* = 4.44; females: 72.05, 76.31, 10.06).

Further tests were conducted to determine if those OSMs dropping out of the BHPS after wave 4 were different in GHQ scores at wave 4 from the panel members. No significant differences were found in the younger panel (males: panel mean 10.27, non-panel mean 10.84, *t* = 1.70; females: 11.69, 11.94, 0.71) while in the older panel, non-panel females had significantly higher scores (males: panel mean 10.26, non-panel mean 11.15, *t* = 1.51; females: 11.36, 12.57, 2.43).

Table 2 presents the comparison of BHPS and HSE GHQ scores for the years 1991-1995 and 1997 for the younger panel. The significance level was set at p<.01 to allow for the larger sample sizes. For males (top panel of the table) there were no significant differences in any of the years and no evidence of any divergence of the panel results from the cross-sectional. For females, significant differences occurred in the results for 1993, 1994, and 1997 along with some evidence that suggested a divergence of the panel results from those of the cross-sections. The direction of the divergence was that the panel results become increasingly higher than the cross-sectional scores, which is not consistent with a social desirability hypothesis.

* * * Table 2 about here * * *

Table 3 presents the same information for the older panel. For these tests, the significance level was set at p<.05 as the sample sizes were smaller than the younger panel. Apart from one year (1991 for males and 1993 for females) in each sequence, no consistent significant differences were found. Also, there was no evidence of divergence for either males or females.

* * * Table 3 about here * * *

The results for younger panel females in Table 2 warranted further investigation. Consequently, the panel and HSE comparison groups were broken down into five ten-year intervals (16-25….56-65 in 1991) and subjected to the same year on year comparisons. Out of the thirty age group/year comparisons, only two were significantly different at the p<.01 level. Both of these were in the 36-45 age group for years 1992 and 1995 when the panel results were higher than the cross-section scores. With the breakdown by age, the results suggested that, for females, the relationship of GHQ score to age was curvilinear, rising from the late 20s to a peak at about 50 and then declining. When the age composition of the panel was compared with the HSE cross-sections, we found that the panel had a higher percentage of females 26-45 and a lower percentage of females 16-25. The over-representation of 26-45 females also coincided with that age group's fastest rate of increase of GHQ scores with age. When taken together, the mostly non-significant differences between panel and cross-section results when broken down by age group and the differing age composition for females in the panel, particularly in the 26-45 age group, suggest that the difference in female GHQ scores observed in Table 2 are a result of panel selection bias. The age composition for males in the panel closely matches those of the HSE comparison groups and in the age breakdown for males only one of the thirty age group/year comparisons was significant at the p<.01 level.

Discussion

These results find no evidence of retest effects in the GHQ-12 for the core sample of the BHPS when compared to the cross-sectional results from the HSEs. They further suggest that, for general population samples, the one-year time period between applications is probably long enough for the respondents not to recall their specific answers, even if they do remember the questions from the year before. The one-year interval used in the Australian Longitudinal Study produced only minor changes (Graetz, 1991) especially when compared to the large retest effects reported by Henderson et al. (1981) with a shorter follow up period.

The differences in GHQ scores for those OSMs in and out of the panel in 1991 and then later in 1994 are not surprising. No differences were found among the younger age group while significant differences were found among the older age group. The non-panel members in the older group were significantly older (average of two years for males and four years for females) and thus more likely to cease being part of the study through illness or death. Up to wave 7, over 70% of the 636 deaths occurred to OSMs over 65 years old.

The finding of a divergence in the panel results from the cross-sectional results for females in the younger panel initially suggested an increased sensitivity to the instrument over time. However, the finer age-graded analysis indicated that this divergence was probably due to the over-representation of females 26-45 and under-representation of females 16-24 in the panel. This age selection bias in the panel then distorted the trajectory of the GHQ scores over the time period of the study. Selection bias in a panel study is to be expected and the BHPS provides a longitudinal weighting scheme to allow for the non-random attrition of the OSMs across waves.

Applying a stricter level of significance to the tests for 1993 to 1995 reduces the strength of evidence for divergence of the panel results. When the HSE comparison groups have a sample size of about 6,000, a level of $p<.01$ may not be strict enough and $t>3.0$ may be more appropriate (Raftery, 1995). This would leave only a significant difference in 1997. However, this would not fundamentally alter the evidence of a divergence, as any divergence through retest effects would expect to be gradual.

The distributions of GHQ-12 scores for males and females with age in both the BHPS panel and HSE cross-sections are similar – rising to middle age and then declining to about 60 and then rising again. These patterns are consistent with those found for an anxiety and depression measure with age in the British National

Household Survey of Psychiatric Morbidity (Bebbington et al., 1998). Although the GHQ-12 does not claim to specifically measure depression, many of its items reflect a depressive or anxious tendency and studies that have factor analysed the items usually name the dominant factor depression and/or anxiety (e.g. Martin, 1999; Schmit et al., 1999; Politi et al., 1994; Graetz, 1991). Therefore, it may be expected to show the same distribution with age as more specific measures of depression and anxiety.

Overall, these results indicate that the GHQ-12 is a consistent instrument over multiple applications with relatively long time periods between applications in general population samples. These properties make it particularly suited for long-term studies that require an indicator of minor psychiatric morbidity.

References

Bebbington PE, Dunn G, Jenkins R, Lewis G, Brugha T, Farrell M, Meltzer H (1998) The influence of age and sex on the prevalence of depressive conditions: report from the National Survey of Psychiatric Morbidity. *Psychological Medicine* 28: 9-19

Bennett N, Dodd T, Flatley J, Freeth S, Bolling K (1995) *Health Survey for England 1993*. HMSO: London

Bowling A (1991) *Measuring health: A review of quality of life measurement*. Milton Keynes: Open University Press

Breeze E, Maidment A, Bennett N, Flatley J, Carey S (1994) *Health Survey for England 1992*. HMSO: London

Buck N (1990) *The British Household Panel Study*. ESRC Data Archive Bulletin 46.

Colhoun H, Prescott-Clarke P, eds (1996) *Health Survey for England 1994.* HMSO: London

Goldberg DP, Williams P (1988) *A user's guide to the general health questionnaire*. Windsor: NFER-Nelson

Graetz B (1991) Multidimensional properties of the general health questionnaire. *Social Psychiatry and Psychiatric Epidemiology* 26: 132-138

Henderson SH, Byrne DG, Duncan-Jones P (1981) *Neurosis and the social environment*. Academic Press Australia: Sydney

Kalton G, Citro CF (1995) Panel surveys: Adding the fourth dimension. *Innovation: The European Journal of Social Sciences* 8: 25-40

Kitamura T, Shima S, Sugawara M, Toda MA (1994) Temporal variation of validity of self-rating questionnaires - repeated use of the general health questionnaire and Zungs self-rating depression scale among women during antenatal and postnatal periods. *Acta Psychiatrica Scandinavica* 90: 446-450

Martin AJ (1999) Assessing the multidimensionality of the 12-item general health questionnaire. *Psychological Reports* 84: 927-935

Prescott-Clarke P, Primatesta P, eds (1998) *Health Survey for England: The health of young people '95-97.* HMSO: London

Ormel J, Koeter, MWJ, van den Brink W (1989) Measuring change with the general health questionnaire: The problem of retest effects. *Social Psychiatry and Psychiatric Epidemiology* 24: 227-232

Politi PL, Piccinelli M, Wilkinson G (1994) Reliability, validity and factor structure of the 12-item General Health Questionnaire among young males in Italy. *Acta Psychiatrica Scandinavica 90: 432-437*

Radovanovic Z, Eric L, Jevremovic I (1988) The effect of retesting on the validity of the general health questionnaire. *Social Psychiatry and Psychiatric Epidemiology* 23: 36-38

Raftery A (1995) Bayesian model selection in social research. *Sociological Methodology* 25: 111-163

Schmidt N, Kruse J, Tress W (1999) Psychometric properties of the General Health Questionnaire (GHQ-12) in a German primary care sample. *Acta Psychiatrica Scandinavica* 100: 462-468

Taylor MF (ed.) with Brice J, Buck N, Prentice-Lane E (1998) *British Household Panel Survey User Manual Volume A: Introduction, Technical Report and Appendices*. Colchester: University of Essex

White A, Nicolaas N, Foster K, Browne F, Carey S (1993) *Health Survey for England 1991*. HMSO: London

Appendix

TABLE 1: Age ranges, sample size and sex composition for Comparison Groups from HSEs 1991-1995 and 1997.

	Age range	N	Males (%)	Females (%)
HSE 1991	16-65	2540	1195 (47.0)	1345 (53.0)
	66+	595	255 (42.9)	340 (57.1)
HSE 1992	17-66	3109	1487 (47.8)	1622 (52.2)
	67+	705	295 (41.8)	410 (58.2)
HSE 1993	18-67	12981	6148 (47.4)	6833 (52.6)
	68+	2456	1028 (41.9)	1428 (58.1)
HSE 1994	19-68	12387	5782 (46.7)	6605 (53.3)
	69+	2355	911 (38.7)	1444 (61.3)
HSE 1995	20-69	12600	5822 (46.2)	6778 (53.8)
	70+	2145	915 (42.7)	1230 (57.3)
HSE 1997	22-71	6711	3086 (46.0)	3625 (54.0)
	72+	885	344 (38.9)	541 (61.1)

TABLE 2: Comparison of Mean GHQ-12 Scores for Younger Panel (16-65 in 1991)

		1991	1992	1993	1994	1995	1996	1997
Males	BHPS	10.01 (4.57)	10.30 (4.67)	10.31 (4.83)	10.27 (4.77)	10.37 (4.83)	10.48 (4.95)	10.38 (4.97)
	HSE	9.85 (4.30)	10.18 (4.51)	10.13 (4.53)	10.09 (4.53)	10.51 (4.84)	-	10.19 (4.55)
	t	1.00	0.78	1.46	1.49	-1.11	-	1.32
Females	BHPS	11.07 (4.83)	11.51 (5.06)	11.60 (5.32)	11.69 (5.44)	11.86 (5.42)	11.92 (5.52)	11.98 (5.78)
	HSE	11.29 (5.13)	11.30 (5.08)	11.24 (5.21)	11.33 (5.00)	11.56 (5.10)	-	11.50 (5.08)
	t	-1.27	1.26	2.80*	2.79*	2.30	-	3.23*

Assumed not to have equal variances. 2-tail tests * p<.01. Standard deviations in brackets.

TABLE 3: Comparison of Mean GHQ-12 Scores for Older Panel (66+ in 1991)

		1991	1992	1993	1994	1995	1996	1997
Males	BHPS	9.95 (3.97)	10.15 (4.22)	10.09 (4.16)	10.26 (4.23)	10.33 (4.30)	10.59 (4.18)	10.71 (4.75)
	HSE	10.75 (4.94)	9.97 (3.84)	10.19 (4.59)	10.01 (4.55)	10.44 (4.79)	-	10.50 (4.77)
	t	-1.96*	0.52	-0.33	0.79	-0.33	-	0.50
Females	BHPS	10.50 (3.92)	10.86 (4.34)	11.54 (4.49)	11.36 (4.21)	11.44 (4.32)	11.46 (4.61)	11.62 (4.71)
	HSE	10.95 (4.62)	11.07 (4.49)	10.92 (4.73)	11.31 (5.03)	11.37 (5.05)	-	11.34 (4.79)
	t	-1.37	-0.64	2.26*	0.17	0.23	-	0.85

Assumed not to have equal variances. 2-tail tests * p<.05. Standard deviations in brackets.