Evidence for the sensitivity of the SF-36 health status measure to inequalities in health: results from the Oxford healthy lifestyles survey

Crispin Jenkinson, Richard Layte, Angela Coulter, Lucie Wright

Abstract

Objectives **– The short form 36 (SF-36) health questionnaire may not be appropriate for population surveys assessing health gain because of the low responsiveness (sensitivity to change) of domains on the measure. An hypothesised health gain of respondents in social class V to that of those in social class I indicated only marginal improvement in self reported health. Subgroup analysis, however, showed that the SF-36 would indicate dramatic changes if the health of social class V could be improved to that of social class I.**

Design **– Postal survey using a questionnaire booklet containing the SF-36 and a number of other items concerned with lifestyles and illness. A letter outlining the purpose of the study was included.**

Setting **– The sample was drawn from family health services authority (FHSA) computerised registers for Berkshire, Buckinghamshire, Northamptonshire, and Oxfordshire.**

Sample **– The questionnaire was sent to 13 042 randomly selected subjects between the ages of 17–65. Altogether 9332 (72%) responded.**

Outcome measures **– Scores for the eight dimensions of the SF-36.**

Statistics **– The sensitivity of the SF-36 was tested by hypothesising that the scores of those in the bottom quartile of the SF-36 scores in class V could be improved to the level of the scores from the bottom quartile of SF-36 scores in class I using the effect size statistic.**

Results **– SF-36 scores for the population at the 25th, 50th, and 75th centiles were provided. Those who reported worse health on each dimension of the SF-36 (ie in the lowest 25% of scores) differ dramatically between social class I and V. Large effect sizes were gained on all but one dimension of the SF-36 when the health of those in the bottom quartile of the SF-36 scores in class V were hypothesised to have improved to the level of the scores from the bottom quartile of SF-36 scores in class I.**

Conclusions **– Analysis of SF-36 data at a population level is inappropriate; subgroup analysis is more appropriate. The data suggest that if it were possible to improve the functioning and wellbeing of those in worst health in class V to those reporting the worst health in class I the improvement would be dramatic. Furthermore, differences between the classes detected by the SF-36 are substantial and more dramatic than might previously have been imagined.**

(*J Epidemiol Community Health* 1996;**50**:377–380)

Health Services Research Unit, University of Oxford, Department of Public Health and Primary Care, Radcliffe Infirmary, Oxford OX2 6HE
C Jenkinson
R Layte

King's Fund Centre for Health Services Development, 126 Albert Street, London NW1 7NF
A Coulter

Academic Department of Psychiatry, Royal Free Hospital School of Medicine, University of London, Pond Street, London NW3 2QG
L Wright

Correspondence to:
Dr C Jenkinson.

Accepted for publication December 1995

A recent paper drew attention to the potential limitations of health and lifestyle surveys which include health status measures in the monitoring of population health over time.[1] In that paper it was suggested that the short form 36 (SF-36) questionnaire might be useful for detecting changes in health in homogenous treatment groups but the variation in responses in a general population might make it inadequate for assessing the impact of health interventions directed at whole communities. The sensitivity of the measure was evaluated by hypothesising a dramatic change in health whereby the health of those in social class V could be improved, by some intervention, to the level of those in social class I. Such an improvement would seem dramatic, yet only small to moderate changes were found on dimensions of the SF-36. This suggests that the instrument would not be an appropriate measure of outcome for community wide interventions.

This paper supports this view, and suggests that for analysis of SF-36 scores at a population level to make any sense it is imperative that the data are analysed at the level of subgroups. The purpose of this paper, therefore, is to demonstrate, using the same dataset as Ziebland,[1] that the SF-36 is sensitive to variations in health between social classes and would detect a change in health of those in social class V to that of those in class I. However, this is only possible if the data is broken down into sub groups at which health care interventions may be targeted, such as those reporting poor health states, and in the lowest social classes where inequalities in health are most marked.[2,3]

Methods

THE SF-36

The original development and validation of the SF-36 has been described extensively in the literature.[4–10] The instrument is a 36 item questionnaire which measures eight multi-item

dimensions: physical functioning (10 items), social functioning (2 items), role limitations due to physical problems (4 items), role limitations due to emotional problems (3 items), mental health (5 items), energy/vitality (4 items), pain (2 items), and general health perception (5 items). There is a further unscaled single item asking respondents about health change over the past year. Minor modifications to the wording of six items on the SF-36 were made to make it acceptable in the British context.[11]

The results reported here are based upon data gained from the Oxford healthy life survey.[12] This was a postal survey in which the SF-36 together with questions on lifestyle and demographics were incorporated into a booklet. A covering letter from the Oxford University Health Services Research Unit was sent with the questionnaire. For those who did not respond to the initial questionnaire, a reminder letter was mailed approximately four to six weeks later. If this elicited no response then another questionnaire and covering letter were sent.

The questionnaire booklet contained, in addition to the SF-36, questions on whether or not the respondent had any long standing illness and had consulted a medical practitioner in the last two weeks because of problems with their own health.

The questionnaire booklet was mailed to 13 042 randomly selected subjects between the ages of 18–65 from family health services authority (FHSA) computerised registers for Berkshire, Buckinghamshire, Northamptonshire, and Oxfordshire.

ANALYSIS STRATEGY

We adopted a strategy similar to that suggested by Ziebland for analysing cross sectional data. Ziebland claims that in the absence of longitudinal data sets the ability of the SF-36 to reflect health gain can be assessed from the analysis of cross sectional surveys. She suggests an analysis assuming that the health of social class V could be improved to match that of social class I using data from the Oxford healthy liefstyles survey. This analysis strategy is followed here. In this paper, the internal reliability of the scales on the SF-36 is assessed separately for the two social classes using the alpha statistic.[13] Data for the eight dimensions of the SF-36 are presented separately for classes V and I in terms of descriptive statistics and 25th, 50th, and 75th centile scores, and effect sizes. The use of 25th, 50th, and 75th centile scores has been recommended by the developers of the measure as a more informative method of gaining a sense of the distribution of scores than group level descriptive statistics in the form of means and SDs.[8] Effect sizes have been advocated as statistically more useful than the less discriminating criteria of statistical significance for the measurement of change over time.[14–16] Floor and ceiling effects[17] (ie those scoring at the top or bottom of the scale) are also documented.

Table 1 Internal reliability (Cronbach's α) for those in social classes V and I

Domain	*α scores for social class V*	*α scores for social class I*
Physical function	0·91 (n=236)	0·84 (n=385)
Social function	0·74 (n=267)	0·79 (n=391)
Role – physical	0·9 (n=258)	0·86 (n=390)
Role – mental	0·83 (n=260)	0·72 (n=391)
Pain	0·84 (n=266)	0·73 (n=387)
Mental health	0·85 (n=258)	0·81 (n=384)
Energy	0·85 (n=258)	0·86 (n=389)
Health perception	0·82 (n=255)	0·74 (n=387)

Results

The questionnaire booklet achieved a response rate of 72% (9332). Altogether 391 (4·2%) respondents were coded as social class I and 267 (2·9%) were coded as social class V using the standard occupational classification.[18–20] The social class composition of the sample has been discussed in previous publications and reflects the characteristics of the general population when compared with 1991 population estimates and the social class distribution in the 1981 census.[12 21] The mean (SD) ages of respondents were 40·6 (10·7) years (range 20–64·7) for social class I and 43·2 (13·0) years (range 18·6–64·7) for social class V. A total of 388 (99·2%) people in class I answered the question on whether or not they had a chronic illness and 116 (30·0%) claimed that they did. Of the class V respondents, 263 (98·5%) answered this question and 75 (28·08%) reported having chronic illness. A χ^2 statistic indicated that these differences were statistically non-significant. However, while reporting of chronic illness was similar in both social classes, use of medical services was higher in social class V. Respondents were asked if they had consulted a doctor on their own behalf in the preceding two weeks. Three hundred and ninety (99·7%) class I respondents answered this question, 52 (13·3%) of whom claimed to have consulted a doctor on their own behalf in the preceding two weeks. Of the 266 (99·6%) respondents in class V who answered this question, 54 (20·4%) claimed to have had such a recent medical consultation. This difference was statistically significant ($\chi^2 = 5·67$; df = 1; $p<0·02$).

Internal reliability of the SF-36 for the entire sample in the Oxford study has been reported previously.[21] Table 1 provides data for the internal reliability of the measure for those in social classes I and V. Internal reliability estimates were high. It has been suggested that an α of 0·50 or above is acceptable,[22] although Nunnally recommends values of 0·70 and above.[23] All domains gained a value higher than these recommended limits. While the results would mediate against using the measure at the individual level, the data are appropriate for group level analysis and comparison, and mirror those gained by the developers of the instrument.[24]

Medians and quartiles for the questionnaire were calculated for respondents in social classes I and V (see tables 2 and 3). Most strikingly at least a 10 point difference can be seen on the 25th centile between the two social groups for all dimensions except for the domains tap-

Table 2 SF-36 scores at the 25th, 50th, and 75th centiles for those in social class I

Social class I (n=384 min)	*25th centile*	*50th centile*	*75th centile*
Physical function	90·00	95·00	100
Social function	88·89	100	100
Role – physical	100	100	100
Role – mental	100	100	100
Pain	77·78	88·89	100
Mental health	68·00	80·00	88·00
Energy	55·00	65·00	80·00
Health perception	67·00	77·00	87·00

Table 3 SF-36 scores at the 25th, 50th, and 75th centiles for those in social class V

Social class V (n=236 min)	*25th centile*	*50th centile*	*75th centile*
Physical function	80·00	95·00	100
Social function	77·78	100	100
Role – physical	75·00	100	100
Role – mental	100	100	100
Pain	66·67	100	100
Mental health	60·00	72·00	88·00
Energy	45·00	60·00	75·00
Health perception	57·00	72·00	87·00

Table 4 Descriptive statistics for social classes V and I for all dimensions on the SF-36 questionnaire together with effect sizes for these data and those reported by Ziebland (see text). Values, mean (SD), 95% confidence intervals, and number

	Social class V	*Social class I*	*Effect size*	*Effect size reported by Ziebland*
Physical function	58·0 (21·4) (52·9–63·1) n=71	80·1 (15·5) (77·1–83·1) n=103	1·03	0·42
Social function	56·2 (20·8) (51·4–61·0) n=76	74·6 (19·8) (71·3–77·9) n=138	0·88	0·25
Role – physical	41·4 (29·3) (34·3–48·5) n=67	82·8 (33·0) (78·8–86·8) n=258	1·41	0·22
Role – mental	79·7 (34·5) (75·5–83·9) n=260	87·3 (26·5) (84·7–89·9) n=391	0·22	0·21
Pain	49·3 (16·1) (45·8–52·8) n=85	66·5 (16·4) (63·3–69·7) n=102	1·07	0·29
Mental health	45·8 (15·9) (42·1–49·5) n=72	56·2 (10·5) (54·1–58·3) n=100	0·65	0·25
Energy	32·8 (10·6) (30·3–35·3) n=71	41·2 (13·1) (38·9–43·3) n=122	0·79	0·41
Health perception	42·3 (14·4) (38·3–45·8) n=69	53·0 (14·2) (50·4–55·6) n=114	0·74	0·23

Table 5 Floor and ceiling effects on the eight dimensions of the SF-36, together with the number completing all items on the dimensions

	Class V			*Class I*		
	% floor	*% ceiling*	*No*	*% floor*	*% ceiling*	*No*
Physical function	0·7	31·5	236	0·3	48·1	385
Social function	0·4	55·4	260	0·3	64·7	391
Role – physical	9·7	71·2	258	4·1	82·6	390
Role – mental	10·9	67·8	267	3·8	77·7	391
Pain	0·7	34·5	267	0·5	46·3	387
Mental health	0·7	3·0	258	0	2·8	384
Energy	0·4	1·9	258	0·5	0·3	389
Health perception	0·4	5·6	255	0	5·6	387

ping "role limitations due to mental health problems" and "mental health".

Table 4 documents descriptive statistics for social classes V and I for the eight dimensions of the SF-36. Assuming an intervention could improve the health status of class V to class I, effect sizes have been calculated to indicate the extent of such an improvement. An effect size of 1·00 is equivalent to a change of 1 SD in the sample. As a bench mark for assessing the relative magnitude of a change, Cohen[25] identified an effect size of 0·20 as small, one of 0·50 as moderate, and one of 0·80 as large. For all dimensions except role limitations due to mental health large effect sizes were gained. These are far greater than those calculated by Ziebland[1] (see also table 4) for the entire sample of respondents in class I compared to those in class V.

Floor and ceiling effects were explored (see table 5). It is worthy of note that a large number of both groups score at the ceiling (ie claim perfect health) for the dimensions of physical functioning, social functioning, role limitations due to physical problems, role limitations due to mental health problems, and pain. Nonetheless social class I gains a greater percentage of respondents reporting perfect health on these particular domains. Furthermore, the apparent equity of role limitations due to mental health problems between the two classes is dispelled with this data. Some 3·8% of social class I fall at the floor while 10·9% fall at the floor in class V.

Discussion

Social class differences in self perceived health are wide and the SF-36 is capable of detecting the differences. However, these differences are most dramatic for those reporting worst health in the different social classes. Thus, the health of the most unwell in class V is substantially worse than the health of the most unwell in class I.

A number of important issues must be borne in mind by potential users of the measure. Firstly, it is, as Ziebland[1] remarks, difficult to ascertain the impact of interventions at a population level using the SF-36.[1] Any changes are likely to appear small in a heterogeneous sample. However, we have suggested a way that may enable researchers to gain better insight into changing patterns of health, notably by analysing those who fall into or onto the lowest quartile score on each dimension of the measure. Nonetheless, while this may provide further evidence for persisting inequalities, it remains unlikely that such data will be of use in any considered attempts to assess what specific health needs are not being addressed. Furthermore, even if the health of those in the lowest social classes improved over time, it remains difficult to ascertain whether the impact of any particular interventions caused such improvements.

We agree with those writers who claim that health status scores at a population level provide too obscure a clue to the exact nature of unmet health care needs and should not, therefore, be used in purchasing decisions.[1 26] However, this paper does support the continued use of measures such as the SF-36 in the monitoring of the health of populations. While such data cannot inform us about the specific health needs of any given population sample, it can provide vital information on the self perceived health of the population. This can be used to support other evidence of inequalities in health and, broken down by regions or districts, may provide data on pockets of particular unmet health needs; further evi-

dence will then be needed to determine what the exact needs of given areas actually are. Information from surveys including the SF-36 will be a useful aid to uncovering the extent of the symptom iceberg[27 28] and, furthermore, highlight geographical and regional differences in health status.[29]

Some caution may be required when interpreting scores. There is evidence of ceiling effects on the SF-36. A previous measure, the Nottingham health profile,[30 31] has been criticised as being too insensitive to lower levels of ill health[32] and the SF-36 has been championed as not suffering from this limitation.[11] Certain dimensions seem to be prone to high levels of perfect scores. It is, therefore, possible that the picture of health provided by some dimensions is somewhat more positive than is really the case. If researchers wish to gain a more sensitive picture of the impact of ill health upon role functioning and social wellbeing it may be necessary to augment the SF-36 with other measures designed to evaluate these specific domains such as the RAND social activities questionnaire[33] and the "work dimension" of the sickness impact profile[34] or functional limitation profile.[35] A number of texts are available which review the most commonly used measures[36 37] while a comprehensive bibliography of available measures is documented regularly in academic journals.[38]

In conclusion, we argue that the SF-36 is an appropriate indicator of ill health in populations which is sensitive to social class variations, and, we suggest, would be able to detect changes in health over time. However, interpretation of the meaning of these differences is not unproblematic and, subsequently, such data are of only limited use in planning health care strategies. While the SF-36 can be used to provide a picture of health inequalities, more detailed information will be required to determine the exact nature of unmet health care needs.

1 Ziebland S. The short form 36 health status questionnaire: clues from the Oxford region's normative data set about its usefulness in measuring health gain in population surveys. *J Epidemiol Community Health* 1995;**49**:102–5.
2 Black D, Morris JN, Smith C, Townsend P. The Black report: inequalities in health. London: DHSS, 1980.
3 Davey Smith G, Bartley M, Blane D. The Black report on socioeconomic inequalities 10 years on. *BMJ* 1990;**301**: 373–7.
4 Ware JE, Brook RH, Williams KN, Stewart AL, Davies-Avery A. *Conceptualisation and measurement of health for adults in the health insurance study. Volume 1: Model of health and methodology.* Santa Monica, CA: RAND Corp, 1980.
5 Stewart AL, Ware JE eds. *Measuring functioning and well being: the medical outcomes study approach.* London: Duke University Press, 1992.
6 Stewart AL, Hays RD, Ware JE. Methods of validating the MOS health measures. In: Stewart AL, Ware JE eds. *Measuring functioning and well being: the medical outcomes study approach.* London: Duke University Press, 1992.
7 McHorney CA, Ware JE, Raczek AE. The MOS 36-item short-form health survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. *Med Care* 1993;**31**:247–63.
8 Ware JE, Snow KK, Kosinski M, Gandek B. *The SF-36 health survey manual and interpretation guide.* Boston: The Health Institute, 1993.
9 Jenkinson C, Wright A, Coulter A. Criterion validity and reliability of the SF-36 in a population sample. *Quality of Life Research* 1994;**3**:7–12.
10 Wright L. The long and the short of it: the development of the SF-36 general health survey. In: Jenkinson C ed. *Measuring health and medical outcomes.* London: UCL Press, 1994.
11 Brazier JE, Harper R, Jones NMB, O'Cathain A, Thomas KJ, Usherwood T, Westlake L. Validating the SF-36 health survey questionnaire: new outcome measure for primary care. *BMJ* 1992;**305**:160–4.
12 Wright L, Harwood D, Coulter A. *Health and lifestyles in the Oxford region.* Oxford: Health Services Research Unit (University of Oxford), 1992.
13 Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrica* 1951;**16**:297–334.
14 Kazis L, Anderson JA, Meenan RF. Effect sizes for interpreting changes in health status. *Med Care* 1989;**27**: S178–S89.
15 Lydick E, Epstein RS. Interpretations of quality of life changes. *Quality of Life Research* 1993;**2**:221–6.
16 Fitzpatrick R, Ziebland S, Jenkinson C, Mowat A, Mowat A. Importance of sensitivity to change as a criterion for selecting health status measures. *Quality in Health Care* 1992;**1**:89–93.
17 Bindman AB, Keane D, Lurie N. Measuring health changes among severely ill patients: the floor phenomenon. *Med Care* 1990;**28**:1142–52.
18 Office of Population Censuses and Surveys. *Standard occupational classification. Vol 1.* London: HMSO, 1990.
19 Office of Population Censuses and Surveys. *Standard occupational classification. Vol 2.* London: HMSO, 1990.
20 Office of Population Censuses and Surveys. *Standard occupational classification. Vol 3.* London: HMSO, 1990.
21 Jenkinson C, Coulter A, Wright L. Short form 36 (SF-36) health survey questionnaire: normative data for adults of working age. *BMJ* 1993;**306**:1427–40.
22 Helmstadter GC. *Principles of psychological measurement.* New York: Appleton-Century-Crofts, 1964.
23 Nunnally JC. *Psychometric theory.* 2nd ed. New York: McGraw Hill, 1978.
24 McHorney CA, Ware JE, Lu JFR, Sherbourne CD. The MOS short-form health survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. *Med Care* 1994;**32**:40–66.
25 Cohen J. *Statistical power analysis for the behavioural sciences.* New York: Academic Press, 1977.
26 Donovan J, Frankel S, Eyles J. Assessing the need for health status measures. *J Epidemiol Community Health* 1993;**47**: 158–62.
27 Hannay DR. *The symptom iceberg: a study of community health.* London: Routledge and Kegan Paul, 1979.
28 Hannay DR. The iceberg of illness and trivial consultations. *J R Coll Gen Pract* 1980;**30**:551–54.
29 Lyons RA, Field H, Littlepage BNC. Measuring health status with the SF-36: the need for regional norms. *J Public Health Med* 1995;**17**:46–50.
30 Hunt S, McKenna S, McEwan P. Measuring health status: a new tool for clinicians and epidemiologists. *J R Coll Gen Pract* 1985;**35**:185–88.
31 Hunt S, McEwan P, McKenna S. *Measuring health status.* London: Croon Helm, 1986.
32 Kind P, Carr-Hill R. The Nottingham Health Profile: a useful tool for epidemiologists? *Soc Sci Med* 1987;**25**: 905–10.
33 Donald CA, Ware JE, Brook RH, Davies-Avery A. *Conceptualisation and measurement of adults in the health insurance study: Vol IV Social health.* Santa Monica: RAND Corp, 1978.
34 Bergner M, Bobbitt RA, Carter WB, *et al.* The sickness impact profile: development and final revision of a health status measure. *Med Care* 1981;**19**:787–805.
35 Patrick D, Peach H. *Disablement in the community.* Oxford: Oxford University Press, 1989.
36 Wilkin D, Hallam L, Doggett M-A. *Measures of need and outcome for primary health care* Rev ed. Oxford: Oxford University Press, 1993.
37 Bowling A. *Measuring disease.* Milton Keynes: Open University Press, 1995.
38 Berzon RA, Simeon GP, Simpson Jr RL, Donnelly MA, Tilson HH. Quality of life bibliography and indexes: 1993 update. *Quality of Life Research* 1995;**4**:53–73.