Journal of Public Health Medicine

Vol. 21, No. 1, pp. 41–45
Printed in Great Britain

SF-36 scores vary by method of administration: implications for study design

Ronan A. Lyons, Kathie Wareham, Malcolm Lucas, David Price, John Williams and Hayley A. Hutchings

Abstract

Background Previous research suggests that people respond differently to health status measures when data are collected by interview or self completion of a questionnaire. The objective of this study was to determine whether SF-36 health status scores differ systematically by method of administration.

Method A randomized cross-over study was carried out on 210 new attenders at general medicine, endocrinology, gastroenterology and urological out-patient departments. The outcome was the difference in SF-36 profiles comparing clinic based interviews with self completion at home by the same subjects.

Results For seven of the eight variables of the SF-36 scores were lower in the self assessment, the differences being statistically significant in four of the eight comparisons. The largest differences were in role limitations due to emotional problems (difference 14.74, 95 per cent confidence interval (CI) 7.76–21.7) and social function (difference 7.21, 95 per cent CI 3.19–11.23).

Conclusions Clinic based interviews systematically exaggerate health status compared with self assessment. The difference is sufficiently large to underestimate the effectiveness of health service interventions when a clinic based pre-intervention and postal self completed follow-up design is used, unless adjustment is made for this systematic bias.

Keywords: SF-36, health status, methodology

Introduction

The MOS 36 item short form 36 (SF-36) is a general health status questionnaire developed for the Rand Corporation's Health Insurance Experiment in the United States,[1,2] and subsequently validated for use in the United Kingdom.[3,4] It measures health related quality of life classified by eight dimensions, namely, physical functioning, social functioning, role limitations due to physical problems, role limitations due to emotional problems, mental health, pain, vitality and general health perception.

One of the useful features of a generic questionnaire is that it can compare health status between different groups of people regardless of the diagnoses. For example, the difference in health status between populations living in geographical regions with different rates of limiting long-term illness has been demonstrated.[5] Another use is to test for differences in health status before and after treatment. In this situation the first questionnaire is often submitted by interview to patients attending hospital whereas the questionnaire following an intervention is submitted by post to save money. However, the mode of administration of a questionnaire may affect the data quality and the resulting scores from them.[6]

One of the earliest studies in the literature comparing personal interviews, mailing, and telephone interviews was carried out by Hochstim in California in the 1960s.[7] He noted that response rates using a single medium were lower for telephone (72 per cent) followed by mail (81 per cent), and personal interview (90 per cent), with missing items being more common in mailed surveys. Further work by Siemiatycki in the 1970s revealed that mail and telephone methods achieved response rates of 70–74 per cent and interviews 84 per cent.[8] However, data on sensitive questions (finance) were missing in 9 per cent of mail, 13 per cent of telephone and 14 per cent of personal interviews, and mail responders were more likely to report depression, anxiety and trouble sleeping than the interviewed group.

Other researchers have noted that elderly people are less likely to respond to postal questionnaires, and that acquiescence bias is more marked in interview surveys and is particularly exaggerated in people with minimal schooling, lower socio-economic status or low mental ability.[9,10]

The stimulus for this research was a previous study of population normative data for the SF-36 carried out by one of

[1]Welsh Combined Centres for Public Health, 41 High Street, Swansea SA1 1LT.

[2]School of Postgraduate Studies in Medical & Health Care, Morriston Hospital, Swansea SA6 6NL.

[3]Morriston Hospital, Swansea SA6 6NL.

Ronan A. Lyons,[1] Consultant/Senior Lecturer in Public Health Medicine

Kathie Wareham,[2] Clinical Research Scientist

Malcolm Lucas,[3] Consultant Surgeon

David Price,[2] Consultant Physician

John Williams,[2] Director

Hayley A. Hutchings,[2] Applied Statistician

Address correspondence to Dr R. Lyons.

the authors (R.A.L.) in 1995.[5] In that study, the SF-36, among other measures, was applied to two random samples of the general population in the same geographical areas, one group being interviewed and the other receiving a postal questionnaire ($n = 1357$). In the interviewed branch of that study mean SF-36 scores were significantly lower in six of the eight comparison, the differences varying by as little as 1.4 points for vitality to 11.2 points for role limitations due to emotional problems. However, the response rate for the interviewed sample was considerably higher (82 vs 58 per cent) and it was not known whether the differences in the SF-36 scores were due to response rate bias or differential response by method of administration. A literature search at that time failed to uncover research papers which dealt with this topic adequately. A study by McHorney *et al*. tested mailed, telephone assisted, and interview administered versions of the SF-36, and reported that health ratings based on the SF-36 were less favourable for mail than telephone respondents.[11] However, the study involved different random groups for each method, and the response rates varied considerably by method so it is not clear to what extent selection and response rate biases could have influenced the results. Because large numbers of researchers were beginning to use the SF-36 as an outcome measure it was felt that this question should be answered urgently.

The purpose of this study was to determine whether people respond differently to the SF-36 health status questionnaire when it is administered by health care professionals in a clinical setting or when used in a self-completed format in the person's own home.

Methods

To test whether the same people complete the SF-36 differently at home by themselves or interviewed at a health care setting, a cross-over design is required. It is also important to test instruments in the type of settings in which it is planned to use such tools. For the purpose of this study three out-patient departments at district general hospitals were chosen. Randomization by method of administration is important to remove the effect of memory of the initial questionnaire on the second questionnaire. A randomized cross-over design was used whereby those agreeing to take part in the study were randomly allocated to either an initial pre-out-patient postal questionnaire followed up by a second interview-assisted questionnaire whilst attending a scheduled out-patient consultation, or an initial interviewer-assisted questionnaire at the out-patient department with a follow-up postal questionnaire.

Postal questionnaires were sent to reach recipients either 10 days before or after an out-patient visit was scheduled. Telephone reminders were used if a questionnaire had not been returned after 4 days following the expected date of receipt of the questionnaire by patients.

New patients attending urological, general medicine and endocrinology, and general medicine and gastroenterology clinics at two district general hospitals were invited to participate in the study. Clinic interviews were carried out by the same group of research nurses at the three clinics. As well as the SF-36 questions, data on age, sex, occupation and diagnosis were collected during the interview.

Social class was assigned from occupation using standard occupational classification codes[12] and grouped into manual (social classes 3 m, 4) and non-manual categories (1, 2, 3 nm).

To have power of 90 per cent to detect a difference of clinical significance (a quarter of a standard deviation) with alpha of 0.05, a sample size of 210 matched responses was required.

Table 1 Characteristics of study participants

Number (% of total)		**Interview, first group 95 (45)**	**Postal, first group 115 (55)**
Age	Mean (range)	59.0 (20–90)	56.1 (17–85)
	SE	1.43	1.54
	Median	60	60
Sex (*n*, %)	Male	50 (53%)	61 (53%)
	Female	45 (47%)	54 (47%)
Medical department (*n*, %)	Surgical	54 (57%)	53 (46%)
	Diabetic	33 (35%)	47 (41%)
	Gastroenterology	8 (8%)	15 (13%)
Social class scale (*n*, %)	SC 1	2 (2%)	5 (4%)
	SC 2	21 (22%)	21 (18%)
	SC 3n	10 (11%)	29 (25%)
	SC 3m	32 (34%)	37 (32%)
	SC 4	18 (19%)	15 (13%)
	SC 5	6 (6%)	4 (4%)
	Missing SC	6 (6%)	4 (4%)

Table 2 Initial and follow-up SF-36 scores by method of administration

	Interview first ($n=95$)			Postal first ($n=115$)			Combined group ($n=210$)
SF-36 variable	Initial	Follow-up	Mean difference interview – postal (95% CI)	Initial	Follow-up	Mean difference interview – postal (95% CI)	Mean difference (95% CI)
Bodily pain	56.0	54.8	1.19 (–3.98, 6.36)	50.6	51.6	1.03 (–2.72, 4.77)	1.10 (–1.98, 4.18)
General health perception	55.2	49.9	5.27 (2.19, 8.35)	45.0	47.3	2.34 (0.9, 4.57)	3.67 (1.81, 5.52)
Mental health	70.9	70.1	0.37 (–2.28, 3.03)	62.1	65.9	3.29 (0.73, 5.85)	2.47 (0.46, 4.48)
Physical function	59.6	57.9	2.72 (–0.83, 6.27)	47.9	53.5	6.37 (3.39, 9.35)	3.79 (1.31, 6.26)
Role limitations due to emotional problems	67.7	57.0	10.99 (1.73, 20.25)	46.3	61.7	14.74(7.76, 21.73)	12.50 (6.94, 18.1)
Role limitations due to physical problems	41.2	30.7	15.66 (7.80, 23.52)	33.9	42.1	5.77 (0.13, 11.41)	10.27 (5.60, 14.9)
Social function	62.1	59.7	2.75 (–2.25, 7.74)	55.8	62.9	7.21 (3.19, 11.23)	5.00 (1.86, 8.14)
Vitality	46.2	47.8	–1.50 (–4.87, 1.87)	37.6	41.2	3.29 (0.26, 6.31)	0.95 (–1.33, 3.22)

Statistical analysis was carried out using the SPSS program. Analysis of variance was used to separate out the effects of postal vs interview and primary vs secondary application of questionnaires (presentation order). One-sample *t*-tests were used to compare before and after SF-36 profiles. Ethical approval was obtained from the West Glamorgan Local Research Ethics Committee.

Results

The study was conducted between 30 August 1996 and 13 January 1997. A total of 380 patients were randomized to achieve 210 completed sets of questionnaires. Among the 170 patients invited to participate but who did not complete the study, 56 cancelled or never attended the clinic, 23 persons refused to participate, 73 agreed to participate but did not return questionnaires, and 18 did not participate for other reasons (e.g. died, blind, mentally impaired, previously entered in study).

Persons randomized to group A (initial questionnaire by post) were more likely to complete both questionnaires (115 vs 95). Table 1 shows the characteristics of study participants. The differences between the groups are small and none of the differences are statistically significant.

Table 2 shows the initial and follow-up SF-36 scores by method of administration, the mean difference in scores and 95 per cent confidence intervals. In those randomized to completing the SF-36 by interview initially, average SF-36 scores are lower in the postal assessment for seven of the eight variables of the SF-36.

In those randomized to completing the SF-36 by post initially, average SF-36 scores are lower in the postal assessment for all eight variables of the SF-36. Ten of the 16 comparisons show statistically significant lower scores for the postal administration.

Differences in postal and interviewed SF-36 scores were not significantly related to gender or manual occupation for any of the eight profiles.

Table 3 shows the results of the analysis of variance for each SF-36 variable by method of allocation (interview vs post) and presentation order. Significant interactions between method of allocation and presentation order occur for role limitations due to emotional and physical problems. Once the effect of the method of allocation is accounted for there is no apparent effect of presentation order. The method of allocation significantly changes the results of SF-36 scores for general health perception, mental health, physical function, role function due to emotional problems and vitality.

Discussion

The results of this study show that there are systematic differences in health ratings for the SF-36 by mode of administration. Postal response ratings are systematically lower than interview-administered ratings.

Table 3 Two-way analysis of variance for each SF-36 variable by method of allocation (interview vs postal) and presentation order (initial vs follow-up)

	Main effects		
	Method of allocation significance of *F*	**Presentation order significance of *F***	**2-way interactions significance of *F***
Bodily pain	0.162	0.979	0.718
General health perception	0.012*	0.561	0.135
Mental health	0.002*	0.473	0.263
Physical function	0.013*	0.546	0.263
Role function emotional	0.048*	0.580	0.002*
Role function physical	0.698	0.250	0.007*
Social function	0.624	0.454	0.138
Vitality	0.002*	0.287	0.661

*p values <0.05.

The importance of these differences lies not so much in whether the two modes of data collection are statistically different but whether these differences are sufficiently large compared with the effect sizes of therapies under evaluation, or the impact of medical conditions on health status. In people with conditions such as diabetes or asthma SF-36 scores are typically 10–20 points lower than in an age- and sex-matched population without such conditions.[13]

Thus the difference in scores by method of administration can equate to 20–50 per cent of the impact of having one of these conditions. As very few interventions are likely to fully restore health status, the potential error in study designs which use an initial clinic based interview with a postal follow-up questionnaire to measure changes in health status is considerable. The largest differences are for role limitations due to emotional and physical problems. However, as there are significant interactions between method of allocation and presentation order for these two variables the magnitude of the difference is hard to interpret. It is also worth noting that a study by Brazier *et al*. showed that these two dimensions of the SF-36 had lower test–retest reliabilities over a two week interval (0.69 and 0.63) than the other dimensions.[3]

It is also possible that the health status of patients responding before the clinical consultation may be different from the health status of patients responding after the consultation – an effect of the consultation. We attempted to remove the effect of the consultation by the random allocation of patients but may not have done so completely as we cannot test for an interaction between the effect of the consultation and method of administration. However, the mean differences in scores by method of administration are similar in magnitude and direction (with the exception of vitality) and so this does not seem to be the case. It is interesting that the differences for bodily pain and vitality are much smaller than those for the other variables. Perhaps this is because those variables contain less sensitive questions than, for example, those contained in the mental health profile, and which might be slightly embarrassing to the interviewee. Support for this hypothesis is to be found in a study by Siemiatycki, who reported lower rates of mental health symptoms in interviews compared with postal surveys.[8]

There is little published information on the comparability of responses to the SF-36 between the administration by face-to-face interviews and by post. A study in the United States by Weinberger *et al*. suggested that there may be large and important differences between the methods of administration, but whether this phenomenon is unique to North American culture is unknown.[14] Inferences from studies using different methods for initial and subsequent data collection may therefore be misleading, with the biases tending to lessen a treatment effect and cause researchers to ignore useful interventions. In the study by Weinberger *et al*., 172 US Veterans tested telephone, face-to-face, and self-administered versions of the SF-36 in a randomized cross-over design. Comparison of face-to-face vs self-administered results was restricted to only 40 individuals, but the direction and size of the differences are similar to those in this study, and strengthen its findings.

Because postal follow-up questionnaires are less expensive than telephone or interview-administered questionaries there is an understandable tendency to use postal questionnaires in health services research or audit. The results of this study do not necessarily mean that postal follow-up studies should cease but that if an initial interview and follow-up design is used the follow-up ratings should be adjusted accordingly. Otherwise such research designs will systematically under-represent the effectiveness of health care interventions.

Acknowledgements

The authors would like to acknowledge the advise and assistance of Professor Ian Russell on study design, and also the research staff at the Clinical Research Unit at the School of Postgraduate Studies in Medicine and Health Care who carried out the data collection.

References

1 Ware JE, Sherbourne CD. The MOS 36 item short-form health survey (SF-36) 1: Conceptual framework and item selection. *Med Care* 1992; **30:** 473–480.

2 Ware EJ, Brock RH, Williams RN, Stewart AL, Davies-Avery A. *Conceptualisation and measurement of health for adults in the health insurance study. Vol. 1. Model of health and methodology*. Santa Monica, CA: Rand Corp., 1980.

3 Brazier JE, Harper R, Jones NMB, *et al*. Validating the SF-36 health survey questionnaire: new outcome measure for primary care. *Br Med J* 1992; **306:** 160–164.

4 Jenkinson C, Coulter A, Wright L. The short form 36 (SF36) health survey questionnaire: normative data for adults of working age. *Br Med J* 1993; **306:** 1437–1440.

5 Lyons RA, Fielder H, Littlepage BNC. Measuring health status with the SF-36. The need for regional norms. *J Publ Hlth Med* 1995; **17**(1): 46–50.

6 Moser CA, Kalton G. *Survey methods in social investigation*, 2nd edn. Aldershot: Gower, 1971.

7 Hochstim JR. A critical comparison of three strategies of collecting data from households. *J Am Stat Assoc* 1967; **62:** 976–989.

8 Siemiatycki J. A comparison of mail, telephone, and home interview strategies for household health surveys. *Am J Publ Hlth* 1979; **69:** 238–245.

9 Herbert R, Bravo G, Korner-Bitensky N, Voyer L. Refusal and information bias association with postal questionnaires and face-to-face interviews in very elderly subjects. *J Clin Epidemiol* 1996; **49:** 373–381.

10 Heal LW, Sigelman CK. Response biases in interviews of individuals with limited mental ability. *J Intellectual Disability Res* 1995; **39:** 331–340.

11 McHorney CA, Kosinski M, Ware JE. Comparison of the costs and quality of norms for the SF-36 health survey collected by mail versus telephone interview: results from a national survey. *Med Care* 1994; **32**(6): 551–567.

12 *Standard Occupational Classification. Volume 2: Social classification and coding methodology*. London: HMSO, 1991.

13 Lyons RA, Lo SV, Littlepage BNC. Comparative health status of patients with 11 common illnesses in Wales. *J Epidemiol Commun Hlth* 1994; **48:** 388–390.

14 Weinberger M, Oddone EZ, Samsa GP, Landsman PB. Are health-related quality of life measures affected by the mode of administration. *J Clin Epidemiol* 1996; **49:** 135–140.

Accepted on 2 September 1998