The effects of question order and response-choice on self-rated health status in the English Longitudinal Study of Ageing (ELSA)

A Bowling, J Windsor

Department of Primary Care and Population Sciences, University College London, London, NW3 2PF, UK; a.bowling@ucl.ac.uk

Accepted 23 February 2007

ABSTRACT

Background: One of the most ubiquitous global health measures is a single self-rated health item. This item may be sensitive to its position in questionnaires and to response-choice wording. The aims of this paper were to investigate the effects of question order and response choice on self-reported health status.

Method: A secondary analysis of wave 1 of the English Longitudinal Study of Ageing (ELSA). Participants were a nationally representative sample of people aged 50 years and over living at home. Over 11 000 respondents were interviewed face-to-face in their homes, and were randomly assigned to one of two versions of a self-rated health item.

Results: The health status item asked after, rather than before, a module of health questions, resulted in more optimal health assessments, although the effect size was small. The version of the health status item with "excellent", rather than "very good" as the first response category resulted in more optimal health assessments, although it had a smaller ceiling effect.

Conclusions: There was support for the insertion of the health status question at the beginning of health questionnaires, as it may be influenced by questions about health and disease if placed at the end, although the effect size was small. Evidence for the version of the item with "excellent", rather than "very good", as the first response choice was more mixed as, although optimism bias appeared higher, the ceiling effects were lower. The smaller ceiling effects for the "excellent" version has important implications for the ability to detect improvements in follow-up studies.

One of the most popular global health measures asks people to rate their health on Likert scales, for example, from "excellent" to "poor", or from "very good" to "very bad". It is a useful summary of diverse aspects of respondents' health;[1] it has been included in many generic health and disease-specific measurement scales,[2] and in surveys worldwide since the 1940s.[3–8] It is a useful addition to both survey and health outcome measures given the substantial body of research that shows that it is significantly and independently associated with medical conditions, health service use, changes in functional status, recovery, mortality and socio-demographics.[9–16] It is likely that the self-rated health question may be sensitive to the response format used, and to its position in the questionnaire.[17]

One principle of questionnaire design is that general questions should be placed before specific questions to minimise bias.[18] In survey research, order effects appear strongest for general or summary questions because they are interpreted in relation to the preceding items.[19] Keller and Ware[20] recommended that generic health status measures should be asked about before specific health and disease questions to remove potential bias from such effects. Not all investigators follow this design, however, with some placing generic health measures after disease-specific questions.[21 22] The effects of such framing in research on health are underinvestigated, but act against methodological guidelines. In theory, then, consideration of areas of life that people have already been questioned about are excluded from subsequent global assessments of life, because respondents judge that they have already answered questions about the former, and thus exclude these from the latter.[23] Therefore, if general health status is asked about after disease-specific questions, then ratings of general health status would be expected to be more favourable because the specific disease or condition has already been considered and excluded from the overall assessment.

Several variations in the wording of the health status item and in its response choices exist. In order to increase the question's discriminative ability, and because of the operation of "social desirability" or "optimism" bias (with most respondents rating their health at the optimal health end of the scale), the developers of the Short Form-36, and others, have added a "very good" category in between the "excellent" and "good" response choices to the question: "In general, would you say your health is…." (five-point Likert response scale);[2] and the Short Form-8 (developed from the Short Form-36) includes an additional "very poor" category at the suboptimal health end of the scale (http://www.sf-36.org/demos/SF-8).

Although there are many investigations of the reliability and validity of the item, fewer studies have investigated potential bias from the item's order in a health questionnaire or from the response choice format used.[24] These issues are important for the accurate interpretation of data, and to make valid comparisons between studies.

The aims of the study were to examine question order and response choice wording effects of the self-rated health item.

MATERIALS AND METHODS

The vehicle for the analyses was the English Longitudinal Study of Ageing (ELSA), which comprises a national random sample of people

born before 1 March 1952 (aged over 50 years), living in private households at baseline. The aim of ELSA is to examine, over time, the relationships between health, economic position and activity, social participation, productivity, networks and support.

The sample was drawn from respondents to the government-funded Health Surveys for England (HSE) in 1998, 1999 and 2001, each of which comprised a separate sample of approximately 16 000 adults. HSE used equal probability sampling methods; each of the HSE samples was designed to be representative of the population in England living in private households. For the HSE sampling, first, postcode sectors were selected from the Postcode Address File for "small users" (ie for private rather than commercial addresses), stratified by health authority and proportion of households in non-manual socioeconomic groups. Then, addresses were then selected systematically from each postcode sector, stratified by health authority and proportion of households in non-manual socioeconomic groups; a specified number of individuals within them were selected for interview.[25]

Households identified in the HSE were listed, and included in the sample frame for ELSA if an adult aged 50 years or over was in residence and had consented to re-contact. Eligible participants were those who were identified as born on or before 29 February 1952 in order to ensure all sample members were aged 50 years and over at the beginning of March 2002, when ELSA interviewing began, lasting over a 12-month period. Seventy per cent of eligible households identified responded to ELSA; and 96% of individuals within these sampled households responded. This equated to an overall individual response rate of 67%. Refusal was the main reason for non-response. No differences were found between the demographic characteristics of the ELSA respondents and the national population using Census data.

ELSA achieved 11 234 interviews (11 030 full, and 204 partial) with eligible sample respondents aged 50 years and over (these formed the core sample). The full ELSA sample also included non-core sample members: 158 proxy interviews, 636 interviews with respondents' partners outside the age range of the study (they were under 50 years of age) and 72 interviews with respondents' partners who had joined the household since the initial sampling, giving 12 100 unweighted core and non-core sample members in total. The non-core sample members were excluded from the analyses presented here. Sample weights were attached to the ELSA dataset to reduce any bias from non-response and enhance confidence in the representativeness of the target population (note: sample weighting can lead to distributions not equalling 100%). This resulted in the weighted sample of the 11 234 core respondents equalling 11 221 observations. These 11 221 observations were randomly assigned to receive one of the two versions of the health status item either at the start or the end of the module of questions about health. They form the basis for the analyses presented here.

The dataset is lodged on the Data Archive at the University of Essex. The full details and explanations of the sampling, response rates, weighting and imputation methods used have been reported elsewhere.[25]

Measures

The ELSA Survey was a face-to-face home interview survey, administered electronically by trained interviewers using laptop computers. The interview schedule included harmonised survey questions that were comparable with those used in UK government and international surveys, including the Survey of Health and Retirement in Europe, and the US Health and Retirement Survey. This was in order to facilitate comparisons between surveys, and to provide data relevant to current policy questions. The development stage of the questionnaire design involved setting up two "expert panels" to agree on the questions, followed by two pilot studies to test the instruments and design.[25] The interview covered mainly finances, and also included two alternative versions of self-assessed health status, physical, psychological and social health and functioning, health behaviours, perceptions of the neighbourhood, and standard socioeconomic status and sociodemographic items.[25–27]

The electronic interview programme randomly assigned respondents to one of two versions of the self-rated health item. One group completed a self-rated health item ("Would you say your health is..") with five response choices from "excellent" to "poor" at the start of the health module. They also completed a self-rated health item ("How is your health in general? Would you say it was...") with five response choices from "very good" to "very bad" at the end of the module. The second group completed the latter at the start of the health module, and the former at the end of it (see Box 1 for the items). This provided an opportunity to examine question order effects (the distribution of responses to the item asked before and after the administration of the module of health questions) and response wording effects of the item.

Statistical analysis

The data were first analysed using descriptive statistics, including frequency distributions and Spearman's ranked correlations. The Wilcoxon rank-sum test for independent groups (also known as the Mann–Whitney U test) and Wilcoxon signed-ranks test for matched pairs were then carried out. Effect sizes were calculated for comparisons between individuals, but not groups (ie these can only be calculated for matched-pairs comparisons).

RESULTS

Characteristics of respondents

Of the 11 221 weighted core sample members, 5204 (46%) were men and 6017 (54%) were women. The mean age was 65.1 years (SD 10.2). The vast majority of the sample were white (97% of men and 98% of women), reflecting the homogeneity of the national population in this older age group. The distributions of respondents by age and sex were similar for both types of

Box 1 The two versions of the health status item, asked at the beginning and end of the health module

Q. 184, 420: How is your health in general? Would you say it was
Very good
Good
Fair
Bad
Or, very bad?
Q. 185, 419: Would you say your health is
Excellent
Very good
Good
Fair
Or, poor?

Table 1 Distribution of responses to the alternative versions of the health status item* asked at the start and end of the module of health questions

	% (n)	% (n)
How is your health in general? Would you say it was	Start of health module† n = 5648	End of health module‡ n = 5560
Very good	26 (1466)	31 (1752)
Good	41 (2304)	40 (2197)
Fair	25 (1417)	22 (1244)
Bad	6 (356)	5 (259)
Very bad	2 (105)	2 (108)
Non-matched Wilcoxon $W = 31.13\times10^6$, $Z = -6.97$; $p<0.001$		
Would you say your health is	Start of health module‡ n = 5559	End of health module† n = 5645
Excellent	14 (764)	12 (686)
Very good	27 (1483)	32 (1773)
Good	30 (1670)	32 (1829)
Fair	20 (1124)	18 (1028)
Poor	9 (518)	6 (329)
Non-matched Wilcoxon $W = 32.12\times10^6$, $Z = -4.66$; $p<0.001$		

*The health status items were randomised for completion either at the beginning or at the end of the module of health questions.
†Group randomised to "very good" version at the start.
‡Group randomised to "excellent" version at the start.

response choice, which were asked at the start of the health module (and at the end).

Effects of question order

Table 1 shows the distribution of responses to the alternative versions of the health status item, when administered at the start and end of the health module. For the version with "very good" as the first response choice ("How is your health in general? Would you say it was..."), the distributions were significantly different at the start and at the end. For example, the proportion of respondents rating their health as "very good" at the start of the health module was 26% (1466), compared with more, 31% (1752), at the end. The findings were similar when analysed by age group (50–64, 65–74, 75–79, 80+ years), apart from those in the 80 years and over age group, in which distributions were not significant.

Table 1 also shows that, for the version of the item "Would you say your health is...", and with "excellent" as the first response choice, slightly fewer respondents rated their health as optimal (as either "excellent" or "very good") at the start of the health module (41%, 2247), compared with the end (44%, 2459). The distributions in the table were significantly different. The findings were similar when analysed by age group (50–64, 65–74, 75–79, 80+ years), apart from those in the age groups 75–79 and 80 years and over, in which distributions were not significant.

The table also indicates that more respondents to both versions of the health status item endorsed the second rather than the first response category; thus more respondents achieved optimal health status ("excellent" or "very good") with the version that included "excellent" as the first response choice. The "excellent" version of the item also showed smaller ceiling effects, however, whether it was asked at the start or at the end of the health module: 14% and 12%, respectively, compared with 26% and 31% with the version with "very good" as the first response choice.

With both forms of response choice, most respondents, except those aged 80 years and over, assessed their health more favourably after, rather than before, completion of the health module. Further analyses examined the possibility that respondents in optimal health, compared with those in suboptimal health, may have perceived their health more positively after considering a battery of medical conditions (ie they were "reassured" about their health). Both respondents who reported a longstanding illness and those who reported a limiting, longstanding illness, however, were significantly more likely than respondents without such conditions to rate their health more favourably after completing the health status item at the end, compared with at the start, of the module. The exception was with respondents aged 75–79 years in relation to the version with "excellent" as the first response choice, and those aged 80 years and over in relation to the version with "very

Table 2 Distribution of responses to the health status items with combined response categories*

Health status (with combined response categories)	Health status asked at start of health module % (n)	Health status asked at end of health module % (n)
Excellent/very good (combined)	33 (3713)	38 (4211)
Good	35 (3975)	36 (4026)
Fair	23 (2541)	20 (2271)
Poor/bad/very bad (combined)	9 (979)	6 (696)
Number of respondents	11 208	11 204
Wilcoxon signed ranks test (matched pairs) $z = -20.15$, $p<0.001$		

*The health status items were randomised for completion either at the beginning or at the end of the module of health questions.

Table 3 Effects of self-rated health item on response choice wording (combined top response choice categories and combined bottom response choice categories)*

Health status (with combined end response categories)	Health status at start of health module "excellent" 1st response choice % (n)	Health status at start of health module "very good" 1st response choice % (n)	Health status at end of health module "excellent" 1st response choice % (n)	Health status at end of health module "very good" 1st response choice % (n)
Excellent/very good (combined)	41 (2247)	26 (1465)	44 (2459)	32 (1752)
Good	30 (1670)	41 (2304)	32 (1829)	39 (2196)
Fair	20 (1122)	25 (1415)	18 (1027)	22 (1242)
Poor/bad/very bad (combined)	9 (518)	8 (460)	6 (329)	7 (367)
Number of respondents	5557	5644	5644	5557
	Non-matched Wilcoxon $W = 30.38 \times 10^6$; Z −11.40; $p<0.001$		Non-matched Wilcoxon $W = 31.02 \times 10^6$; Z −11.38; $p<0.001$	

*Unrelated groups in each category.

good'' as the first response choice. (All detailed tables available on request from the authors.) Missing responses to the health status items were small (<20).

In order to make ''matched pairs'' comparisons between responses to the health status item placed before and after the health module, the different response categories at the optimal end (''excellent'' and ''very good'') of both scales were combined, as were the response categories at the suboptimal (''poor'', ''bad'' and ''very bad'') end of the two scales. Table 2 shows that when ''before'' and ''after'' responses were compared, the differences were highly significant. For example, the proportions of all respondents who reported optimal health (''excellent'' or ''very good'') were 33% (3713) at the start, and 38% (4211) at the end. In all age groups, individuals' recoded responses to the combined scale at the end of the health module indicated better health than did their recoded responses at the start. This significant difference ($p<0.001$) was obtained for all the different age ranges compared (50–64, 65–74, 75–79, 80 years and over). The effect size was, however, small (0.1187/0.95 = 0.1249).

Effects of response choice wording

The distributions of the two health status questions (''excellent'' or ''very good'' as the first response choice), asked at the start of the health module, were compared using responses recoded to the combined scale. The distributions at the end of the health module were also compared. Table 3 shows the results of two tests. The first test compared those who answered the ''excellent'' version of the item at the start of the health module with those who answered the ''very good'' version at the start. The second test compared those who answered the ''excellent'' version at the end of the health module with those who answered the ''very good'' version at the end.

The results indicate that respondents who completed the ''excellent'' rather than the ''very good'' version of the health status item at the start of the health module had more optimal health: 41% (2247) versus 26% (1465), respectively, rated their health as ''excellent/very good''. The findings were similar when analysed by age group (50<65, 65<75, 75<80, 80 years and over). These differences were confirmed when the distributions to these alternative response choices were compared at the end of the module (see table 3). Comparisons of the results at the suboptimal ends of the scale (''poor/bad/very bad'') were less conclusive, possibly reflecting the relatively small proportions of respondents who endorsed these categories. The health status version that provided more response choices for suboptimal health (''bad'' or ''very bad'' rather than just ''poor'') did not elicit more negative assessments of health overall.

DISCUSSION

This study found that the commonly used subjective health status item is influenced by question order. When the item was asked after, rather than before, the module of health questions in ELSA, it resulted in more favourable self-assessments of health. This suggests that the item may be influenced by questions about health and disease if placed at the end. Although the effect size was small, the results are consistent with the principles of questionnaire design. For example, because specific diseases and conditions had already been considered, the latter were excluded from the subsequent overall assessment. There was no effect of question order on item response.

The analyses also found that the health status version with ''excellent'' as the first response choice (''Would you say your health is...'') resulted in significantly more optimal health assessments than the version starting with ''very good'' (''How is your health in general? Would you say it was...?''), suggesting that the former was more vulnerable to optimism bias. The number of missing cases was small for both versions of health

What this paper adds

- Rarely have investigators compared the responses of people who have been randomly assigned between alternative versions of the internationally used health status item. This paper makes a unique contribution to the literature on health status by analysing such data.
- Two versions of the ubiquitous self-rated health status item were compared in an English population sample aged 50 years and over.
- The results support the insertion of the health status question at the beginning of interview questionnaires to minimise bias, although the effect size was small.
- Evidence for the version of the item with ''excellent'' rather than ''very good'' as the first response choice was more mixed as, although optimism bias appeared higher, the ceiling effects were lower.
- The smaller ceiling effects for the ''excellent'' version has important implications for the ability to detect improvements in follow-up studies.
- The research needs replicating with different populations and in different countries.

Policy implications

The use of health status measures with sound psychometric properties is essential for the validity of research, the accuracy of observations made, and for the soundness of health policy decisions based on research.

status. It is unlikely that the slight variation in the actual question wording was influential, although this could not be assessed independently of the different response formats.

More respondents endorsed the second rather than the first of the optimal health response categories (ie "good" rather than "very good", or "very good" rather than "excellent". The reasons for this are unknown, although the methodological literature indicates that people prefer to appear "average" preferring to endorse middle-range response categories rather than extreme values.[18] The version of the health status item with "excellent" as the first response category resulted in smaller ceiling effects. This finding is important because investigators in trials and longitudinal surveys need to avoid ceiling and floor effects in their selected measures, in order to be able to detect changes at follow-up assessment. The health status version that provided more response choices for suboptimal health ("bad" or "or, very bad" rather than just "poor") did not elicit more negative assessments of health overall, and does not therefore appear to address optimism, or social desirability bias, more satisfactorily.

It has been argued that it is surprising that the question works so well,[28] especially as Crossley and Kennedy[29] also reported that a substantial minority of respondents changed their response to a health status item when it was included twice in a questionnaire. We were unable to assess the biasing effects of including two health status items in the questionnaire as there was no control group for this. This is worthy of future investigation.

The study was based on a face-to-face interview survey of a population sample aged 50 years and over living in the community in England. The strength of the survey was its large sample size, and with a relatively good response rate (67%). Generalisations from the data are accordingly limited to comparable populations. Also, the response rate still leaves a third of people who may differ in some unknown way from the respondents. It should also be pointed out that the sample was taken from samples of people who responded to earlier health surveys, and who consented to re-contact. The weakness of this approach is the potential cumulative sample bias. The findings are also specific to face-to-face interview formats of questionnaire administration. It is well known that the method of questionnaire administration can also affect the type of response as well as response rates, limiting comparisons of data obtained from different forms of questionnaire administration.[18] These research findings need testing in other populations, in other countries, and using other types of questionnaire administration. A small number of investigators has examined the structure of health questionnaires on responses, and reported inconclusive results; most studies have also been limited to postal surveys.[30 31] The use of health status measures with sound psychometric properties is essential for the validity of research, the accuracy of observations made, and for the soundness of health policy decisions based on research.

Acknowledgements: The authors would like to acknowledge M. Marmot *et al*, who are the original data creators, depositors and copyright owners of ELSA, the funders of the Data Collections and the UK Data Archive. The bibliographic citation for the electronic data collection is: Marmot M, *et al*, English Longitudinal Study of Ageing: Wave 1, 2002–2003 [computer file]. 3rd Edition. Colchester, Essex: UK Data Archive [distributor], September 2005. SN: 5050. The original data creators, depositors or copyright holders, the funders of the Data Collections and the UK Data Archive bear no responsibility for their further analysis or interpretation.

Competing interests: None declared.
Contributors: AB conceived the idea for the study, and with JW designed the framework for the analyses. JW undertook the statistical modelling and interpreted the statistical findings jointly with AB. AB wrote the paper.

REFERENCES

1. **Bowling A.** Just one question: if one question works why ask several? *J Epidemiol Community Health* 2005;**59**:342–5.
2. **Ware JE,** Snow KK, Kosinski M, *et al*. *SF-36 health survey*. Boston MA: New England Medical Centre, 1993.
3. **Ware JE,** Davies-Avery A, Donald C. *Conceptualisation and measurement of health for adults in the health insurance study*: Vol. V, general health perceptions. R-1987/5-HEW.Santa Monica, CA: Rand, 1978.
4. **Brook RH,** Ware JE, Davies-Avery A, *et al*. *Conceptualisation and measurement of health for adults in the health insurance study*: Vol. VIII, overview. R-1987/88-HEW.Santa Monica, CA: Rand, 1979.
5. **Stewart AL,** Ware JE, editors. *Measuring functioning and well-being. The medical outcomes study approach*. Durham: Duke University Press, 1992.
6. **Thompson WE,** Streib GF. Situational determinants: health and economic deprivation in retirement. *J Soc Issues* 1958;**14**:18–45.
7. **Schnore LF,** Cowhig JD. Some correlates of reported health in metropolitan centers. *Soc Problems* 1959;7:218–26.
8. **Cartwright A.** *Health surveys in practice and in potential*. London: Kings Fund, 1983.
9. **Kaplan GA,** Camacho T. Perceived health and mortality: a nine-year follow-up of the Human Population Laboratory Cohort. *Am J Epidemiol* 1983;**117**:292–8.
10. **Goldstein MS,** Siegel JM, Boyer R. Predicting changes in perceived health status. *Am J Publ Health* 1984;**74**:611–15.
11. **Schoenfeld DE,** Malmrose LC, Blazer DG, *et al*. Self-rated health and mortality in the high-functioning elderly – a closer look at healthy individuals: MacArthur field study of successful ageing. *J Gerontol (M)* 1994;**49**:109–15.
12. **Idler EI,** Kasl SV. Self-ratings of health: do they also predict change in functional ability? *J Gerontol (B)* 1995;**50**:S344–53.
13. **Greiner PA,** Snowdon DA, Greiner LH. Self-rated function, self-rated health, and postmortem evidence of brain infarcts: findings from the Nun study. *J Gerontol (B)* 1999;**54**:S219–22.
14. **Bierman, BS, Bubolz TA, Elliott A**. How well does a single question about health predict the financial health of Medicare managed care plans? *Effect Clin Pract* 1999;**2**:56–62.
15. **Siegel M,** Bradley EH, Kasl SV. Self-rated life expectancy as a predictor of mortality: evidence from the HRS and AHEAD surveys. *Gerontology* 2003;**49**:265–71.
16. **Spiers N,** Jagger C, Clarke M, *et al*. Are gender differences in the relationship between self-rated health status and mortality enduring? Results from three birth cohorts in Melton Mowbray, United Kingdom. *Gerontologist* 2003;**43**:406–11.
17. **Bowling A,** Bond M, Jenkinson C, *et al*. Short form-36 (SF-36) health survey questionnaire: which normative data should be used? Comparisons between the norms provided by the Omnibus Survey in Britain, The Health Survey for England and the Oxford Health and Lifestyle Survey. *J Public Health Med* 1999;**21**:255–70.
18. **Bowling A.** *Research methods in health. Investigating health and health services*. 2nd edn. Buckingham: Open University Press, 2002.
19. **Dooley D.** *Social research methods*. New Jersey: Englewood Cliffs, 1995.
20. **Keller SD,**Ware JE. Questions and answers about SF-36 and SF-12. *Medical Outcomes Trust Bulletin* 1996;**4**:3.
21. **Prescott-Clarke P,** Primatesta P, editors. *Health survey for England, 1996*. Vols 1 and 2. London: The Stationery Office, 1998.
22. **Coates A,** Porzsolt F, Osoba D. Quality of life in oncology practice: a prognostic value of EORTC QLQ-C30 scores in patients with advanced malignancy. *Eur J Cancer* 1997;**33**:1025–30.
23. **Sudman S,** Bradburn NM. *Asking questions*. New York: Jossey Bass, 1983.
24. **Eriksson I,** Undén AL, Elofsson S. Self-rated health. Comparisons between three different measures. Results from a population study. *Int J Epidemiol* 2001;**30**:326–33.
25. **Marmot M,** Banks J, Blundell R, *et al*. *Health, wealth and lifestyles of the older population in England. The 2002 English longitudinal study of ageing*. London: Institute of Fiscal Studies, 2003.
26. **Goldberg DP,** Williams P. *A user's guide to the General Health Questionnaire*. Windsor: NFER-Nelson, 1988.
27. **Hyde M,** Wiggins RD, Higgs P, *et al*. A measure of quality of life in early old age: the theory, development and properties of a needs satisfaction model (CASP-19). *Ageing Ment Health* 2003;7:186–94.
28. **Fayers PM,** Sprangers MAG. Understanding self-rated health. *Lancet* 2002;**359**:187–8.
29. **Crossley TF,** Kennedy S. The reliability of self-assessed health status. *J Health Econ* 2002;**21**:643–58.
30. **Dunn KM,** Jordan K, Croft PR. Does questionnaire structure influence response in postal surveys? *J Clin Epidemiol* 2003;**56**:10–16.
31. **McColl E,** Eccles MP, Rousseau NS, *et al*. From the generic to the condition-specific? Instrument order effects in quality of life assessment. *Med Care* 2003;**41**:777–90.