Single item measures

Just one question: If one question works, why ask several?

Ann Bowling

While shorter instruments are more limited than longer measures, they have obvious benefits for both research and policy in terms of reduced burden and costs, and ease of interpretation.

A frequently asked question by clinical investigators is why they should use a lengthy, multi-item measurement scale to assess patients' perceptions of their health, or quality of life, when there is evidence that a measure containing a single, global question is likely to suffice. Researchers may not wish to use lengthy scales because their core questionnaires are already long, the patient group of interest is ill or frail, they wish to minimise the burden on the patient and on the research team, or they simply want a ''snap shot'' of a topic rather than comprehensive coverage. In such circumstances, single questions have the obvious advantage of brevity, of making fewer demands than multi-item measures on respondents and researchers. Single, global questions have long been used in population surveys to measure health status, quality of life (QoL), and health related quality of life (HRQoL). The two most popular single global health items are self rated health status and self reported limiting, longstanding illness.

SELF RATED HEALTH STATUS

The classic self rated health status item consists of asking respondents to rate their health as ''excellent, good, fair, or poor''. Variations of this question have been used in surveys worldwide. Literature reviews on the conceptualisation and measurement of health published by Rand in the USA[1 2] and an overview by Stewart and Ware[3] reported citations of the self rated health item as early as early as the 1950s. For example, a version appeared in a US study of occupational retirement[4] and in the US Federal Civil Defense Administration Survey, both in the 1950s.[5] And a question asking people to rate their general health, followed by a broad question about ill health (including longstanding complaints) was also included in the British government Surveys of Sickness conducted between 1943 and 1952.[6]

Interest in using broad, subjective health items dates from the mid-20th century, and stemmed from the realisation that mortality was too insensitive to use as a health care outcome indicator in developed countries, that health has physical, mental, social and spiritual dimensions, and that patients' perspectives of their health and health outcomes should be assessed. This was given impetus by the World Health Organisation's abstract conceptualisation of health in its 1946 constitution as ''a state of complete physical, mental and social wellbeing, and not merely the absence of disease and infirmity'',[7] and by subsequent investigations of lay definitions of health, and variations in illness behaviour. Survey researchers found that single items measuring subjective health and wellbeing did not necessarily correlate with medical diagnoses, but the former were held to have greater validity in certain situations (for example, when predicting help seeking behaviour and health service use). Thus, while in the first half of the 20th century the focus of health measurement was often limited to the presence or absence of negative health states and functioning, during the last half of the century there was a shift in focus. There was a trend in survey research towards using the single global health status question to integrate the different dimensions of health emphasised in the WHO definition. Consistent with this changing emphasis, the British government's General Household Survey (GHS) included a version of the single health status item asking respondents to rate their health from 1977 onwards, after deciding to broaden its emphasis from use of services in relation to chronic and acute illnesses, and check lists of symptoms, and towards subjective perceptions of health.[6] The question, or a similar variant, was included in the US National Health Interview Survey (http://www.cdc.gov/nchs/nhis.htm) and US National Health and Nutrition Examination Survey (http://www.cdc.gov/nchs/nhanes.htm). Most OECD countries now conduct regular population health interview surveys that include this well known single item (http://www.oech.org/publications). It has also been used with satisfactory levels of validity and reliability in the developing world (for example, Tanzania).[8]

The item was used in the Rand health insurance experiment and medical outcomes study,[3] and now forms part of the general health perceptions dimension in the most widely used multi-item, multi-dimensional health status measure of all, which was developed from the initial Rand measures—the short form-36 (SF-36) (in both the Rand (http://www.rand.org) and QualityMetric (http://www.sf36.org) versions. In the late 1970s, to increase the question's discriminative ability, and because of the operation of ''social desirability'' or ''optimism'' bias (leading to most respondents to rate their health at the positive end of the scale), the developers of the SF-36 and others added a ''very good'' category in between the ''excellent'' and ''good'' response choices; the short form-8 (developed from the SF-36) also includes a ''very poor'' category at the other end of the scale (http://www.sf-36.org/demos/SF-8). The health status item is popular in social gerontology where the tradition has been to ask respondents to rate their health in relation to their age. This prevents older respondents from assessing their health with reference to younger age groups and thereby perceiving it to be suboptimal.

A substantial body of international research has reported the item to be significantly and independently associated with specific health problems, use of health services, changes in functional status, recovery from episodes of ill health, mortality, and sociodemographic characteristics of respondents.[9–17] It is judged to be appropriate for use in population surveys. Investigators of the MacArthur field study of successful aging in the USA, for example, reported that self rated health (poor/bad ratings of health compared with excellent ratings) was a strong and significant predictor of mortality in the general sample, as well as in controlled analyses when the sample was divided into in healthy and less healthy cohort samples.[17] The question has been shown to discriminate successfully between people in different ethnic groups in Britain (http://www.archive.officialdocuments.co.uk/documents/doh/survey99/hse99),

Abbreviations: QoL, quality of life; HRQoL, health related quality of life; GHS, General Household Survey; VAS, visual analogue scale

between indigenous and non-indigenous Australians (http://www.abs.gov.au/ausstats/abs@nsf), and between Maori and other New Zealand subgroups (http://www.moh.govt.nz/moh.nsf), although it is unknown whether differences also reflect cultural variations in perceptions and reporting.

However, variations between surveys and nations in the wording of the item, and in the number of response categories, do limit comparative analyses and interpretations. Analysis of data from the Australian National Health Survey has shown that it does have some response instability when repeated in the same questionnaire (before and after other questions about health), although this might also reflect the biasing effect of question order.[18] And interpretation of the item at an individual level varies, depending on the referent being used by the respondent. Some people refer to specific health problems and others refer to general physical functioning when replying to the question.[19] Other research using anchoring vignettes (fixed descriptions of each response choice level, to increase consistency of respondents' interpretations of them), has found that their use provides a powerful tool for adjusting for the influence of varying expectations on self ratings of health.[20] This can improve comparison of results (for example, older and younger people with the same level of health might rank themselves differently on a health status scale because of varying expectations of health and ability by age).

OTHER POPULAR SINGLE ITEMS

The second most popularly used single item measures disability by asking respondents if they have any ''longstanding illness, disability or infirmity''. Respondents who report positively are usually asked if this limits their activities in any way. The British GHS has included this item since 1972. Although the prevalence of longstanding illness has been shown to increase over the past three decades of the GHS, the pattern has fluctuated.[21] Thus, while the question has been shown to be associated with health service use, mortality, other indicators of functioning and health, age, socioeconomic status, as well as self rated health,[22] the question has posed an enigma for researchers when comparing international data and data over time. A review of the use of the question (and variations of it) reported that it produced estimates that were sensitive to question wording and question order effects, to the mode of data collection (for example, interviewer compared with self administered questionnaires), to the survey process (for example, the collection of data by proxy) and the sponsorship or contextual effects of the survey.[23] It was concluded that estimates of disability using such subjective single item questions were less stable for people who were above, than below, state pension age; and unless surveys that use the same single item instrument follow identical survey procedures, the interpretability of any evidence of change over time is seriously compromised. If single item questions are to be used, then attention to clear, simple wording at their design stage is obviously essential.

The visual analogue scale (VAS) is another frequently used single item technique. The method uses lines, the lengths of which are taken to denote the continuum of some experience such as tiredness, pain, nausea, or anxiety. The lines are usually horizontal, 10 cm in length, with stops (''anchors'') at right angles to the line at both extremes, representing the limits of the experience being measured (for example, ''severe pain'' to ''no pain at all''). The respondent places a cross on the line to indicate their state. A quality of life VAS (often called a ''QoL uniscale'') is in widespread use, in which the respondent places a cross on a horizontal line to indicate their quality of life during a specified time period (anchored at each end from ''lowest quality'' to ''highest quality''). There are many references in the literature to the high levels of reliability, validity, and sensitivity of this simple VAS technique, including its ability to discriminate between healthy and sick people, its sensitivity to the stages of the disease progress, and ability to predict mortality. Research with cancer patients has also shown that a single item QoL VAS has good to excellent levels of reliability and validity compared with multi-item measures.[24]

SINGLE COMPARED WITH MULTI-ITEM MEASURES

Single item measures can be used alongside multi-dimensional measures, and are useful as broad summary ratings of diverse aspects of respondents' health, QoL, and HRQoL, especially where respondents might have improved on one domain (for example, physical functioning) but not on another (for example, mental functioning). They are also generally accepted as useful in the assessment of health transitions (for example, self assessments of health as ''better, same, or worse''). It has been proposed that concepts such as health status, QoL and HRQoL, when used as outcome variables, are more appropriately measured with a global single item.[25] This is because multi-domain measures confound the dimensionality of these concepts with the multiplicity of their causal sources. Thus, in order that predictor and component variables can be separated, such concepts need to be considered as unidimensional, but with multiple causes. The unidimensional indicator is then logically the dependent variable in analyses, and the predictor variables include the range of pertinent multi-dimensional scale variables (for example, social, psychological, functional ability, etc).

While the single item question can provide valuable information, it has the advantage of simplicity, and can be reliable and valid, it is at the expense of detail. More information may be required on different dimensions of health, QoL or HRQoL, than a single item can provide. Classic measurement theory holds that single items are at a relative disadvantage to multi-item measures because more items produce replies that are more consistent and less prone to distortion from sociopsychological biases, and this enables the random error of the measure to be cancelled out. Hence they are more stable, reliable, and precise.

The careful development work on health status batteries at Rand in the USA has shown that a well constructed multi-item scale (even with just 5–10 items) is more sensitive to changes in patients' condition over time than any single item measure.[26–28] In addition, multi-item measures can provide a complete profile of multidimensional phenomenon, and can yield information on changes within the individual dimensions measured by the scale (for example, physical functioning, psychological health), although at the cost of increased burden and the risk of asking irrelevant questions. Scales may be preferred to single items because their multiple items are suitable for statistical calculations using summed and weighted scores (for example, pain might be given twice the weight of mobility in the scale score, if it is judged to be twice as important). On the other hand, there is a body of literature in psychology that shows there is little to be gained by complex weightings over simple summated scoring methods.[29]

PSYCHOMETRIC THEORY

Few of the initially developed measures of health status or HRQoL were based on established methods of scale construction, although methods of scaling had been developed in the early 20th century (for example, Thurstone, Likert, and Guttman techniques of scaling), stemming, in particular, from the development of occupational and intelligence

testing, and the scientific principles of measurement established by mathematical psychologists during the mid-20th century. These led to the establishment of rigorous methods of psychometric evaluation.[30 31] Psychometric theory dictates that when a concept cannot be measured directly (for example, health status, QoL, HRQoL), a series of questions that tap different aspects of the same concept need to be asked. Items can then be reduced, using specific statistical methods, to form a scale of the domain of interest, and the resulting scale tested to ensure that it measures the phenomenon of interest consistently (reliability), that it is measuring what it purports to measure (validity), and is responsive to relevant changes over time. The satisfaction of these conditions is most probable when the resulting instrument contains several items to measure the concept of interest to permit testing for internal consistency and to minimise random measurement error. Although developed much earlier, these standards for measurement and scaling were little used in the health field until the 1970s. Thereafter, during the 1980s and 1990s, emerging patient based health status and HRQoL measures were notable for their length, sometimes containing well over 100 questions, their length being dictated by the rigours of the theories and methods of scaling and psychometrics. More recently, burdened by the length of such measures, investigators have welcomed the development of briefer measures. Hence there has been a proliferation of increasingly short versions of existing measurement instruments, and more efficient summary measurement scales for use in the burgeoning health outcomes sphere.[28]

One of the earliest and most extensive applications of psychometric theory and methods in the health measurement field, began in the 1970s with the refinement of health status measures for Rand's health insurance study and medical outcomes study.[3] One aim of the latter was to construct the best possible, and most efficient, scales for measuring a wide range of functioning and wellbeing. The Rand investigators also realised that new standards of measurement were needed because while traditional testing showed that longer measurement scales were more reliable and valid than shorter scales, they needed to consider respondent burden and the costs of data collection for their large scale studies. They saw the need to compromise between traditionally defined standards of psychometric excellence and newly identified standards of feasibility and practicality; and took as their starting point the issues of which concepts should be measured and how much measurement would be enough for the intended purpose. They attempted to achieve reductions in respondent burden without sacrificing measurement precision below a critical level, and their achievements are apparent when contrasting the number of items in the measuring instruments used in their health insurance experiment with the smaller number of items for their later medical outcomes study (for example, 25 items compared with 10 items to measure physical functioning). These methodological developments have continued, and probably the most well known example of these is the development of the short form-12 (12 items) and the short form-8, both derived from the short form-36 health status questionnaire, as well as the development of summary measures of physical and mental health (see http://www.sf-36.org) and http://www.rand.org).

While the shorter versions of these short form scales are inevitably less sensitive than the full versions, their careful and thorough psychometric development and calibration, based on the most powerful items from the parent instruments, has led to their retaining a high degree of accuracy, and hence their increasing popularity in research on clinical outcomes and population health. The longer SF-36 contains several questions to measure each of the eight dimensions that it includes (physical and social functioning, physical and emotional role limitations, mental health, energy/vitality, pain, and general health perceptions), but the SF-8 derived from it contains just one single item to measure each of these same eight domains. Moreover, the health perceptions item is a variation of the long used health status question: ''Overall, how would you rate your health in the past four weeks? Excellent, very good, good, fair, poor, or very poor?'' Thus the robustness of this item, which has been used, with small variations, as a single item measure of health status in population surveys for over half a century, has at last achieved authoritative acknowledgement.

In conclusion, with the use of more advanced statistical and psychometric techniques, and with awareness of the need to balance psychometric acceptability with practicality, scale developers have responded positively to the frequently asked question: ''If one question works, why ask several?'' Investigators now have an evidence base to guide their selection of longer or shorter multi-dimensional scales and/or single item measures, depending on the purpose and needs of the study. While shorter instruments are more limited than longer measures, they have obvious benefits for both research and policy in terms of reduced burden and costs, and ease of interpretation.

ACKNOWLEDGEMENTS

I would like to thank Professors Emily Grundy, Cathy Sherbourne, and John Ware for helpful information on the history of single item measures.

J Epidemiol Community Health 2005;**59**:342–345.
doi: 10.1136/jech.2004.021204

Correspondence to: Professor A Bowling, Department of Primary Care and Population Sciences, University College London, Royal Free Campus, Rowland Hill Street, London NW3 2PF, UK; a.bowling@ucl.ac.uk

REFERENCES

1 **Ware JE**, Davies-Avery A, Donald C. *Conceptualisation and measurement of health for adults in the health insurance study*. Vol V. *General health perceptions*. Santa Monica, CA: Rand, 1978.
2 **Brook RH**, Ware JE, Davies-Avery A, *et al*. *Conceptualisation and measurement of health for adults in the health insurance study*. Vol VIII. *Overview*. Santa Monica, CA: Rand, 1979.
3 **Stewart AL**, Ware JE, eds. Measuring functioning and well-being. *The medical outcomes study approach*. Durham: Duke University Press, 1992.
4 **Thompson WE**, Streib GF. Situational determinants: health and economic deprivation in retirement. *Journal of Social Issues* 1958;**14**:18–45.
5 **Schnore LF**, Cowhig JD. Some correlates of reported health in metropolitan centers. *Social Problems* 1959;**7**:218–26.
6 **Cartwright A**. *Health surveys in practice and in potential*. London: Kings Fund, 1983.
7 **World Health Organisation**. *Constitution of the World Health Organisation*. Geneva: World Health Organisation, 1948.
8 **Smide B**, Whiting D, Mugusi F, *et al*. Self-perceived health in urban diabetic patients in Tanzania. *East Afr Med J* 1999;**76**:67–70.
9 **National Heart and Lung Institute**. Report of a task group on cardiac rehabilitation. Proceedings of the Heart and Lung Institute Working conference on health behaviour, Bethesda, MD, US Department of Health, Education and Welfare, 1976.
10 **Singer E**, Garfinkel R, Cohen SM, *et al*. Mortality and mental health: evidence from the midtown Manhattan re-study. *Soc Sci Med* 1976;**10**:517–21.
11 **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.
12 **Goldstein MS**, Siegel JM, Boyer R. Predicting changes in perceived health status. *Am J Public Health* 1984;**74**:611–15.
13 **Idler EI**, Kasl SV. Self-ratings of health: do they also predict change in functional ability? *J Gerontol (B)* 1995;**50**:S344–53.
14 **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.
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 **Bierman BS**, Bubolz TA, Elliott A. How well does a single question about health predict the financial health of Medicare managed care plans? *Effective Clinical Practice* 1999;**2**:56–62.
17 **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* 1994;**49**:M109–15.

18 **Crossley TF**, Kennedy S. *The stability of self-assessed health status. Social and economic dimensions of an aging population.* (SEDAP research paper no 26). Canberra, ACT: SEDAP, Australian National University, 2000.
19 **Meurer LN**, Layde PM, Guse CE. Self-rated health status: a new vital sign for primary care? *Wis Med J* 2001;**100**:35–9.
20 **Salomon JA**, Tandon A, Murray CJL. Comparability of self-rated health: cross sectional multi-country survey using anchoring vignettes. *BMJ* 2004;**328**:258.
21 **Walker**, Maher J, Voulthard M, *et al. Living in Britain. Results from the 2000 General Household Survey.* London: The Stationery Office, 2001.
22 **Manor O**, Matthews S, Power C. Self-rated health and limiting longstanding illness: inter-relationships with morbidity in early adulthood. *Int J Epidemiol* 2001;**30**:600–7.
23 **Bajekal M**, Harries T, Breman R, *et al. Review of disability estimates and definitions.* (In house report no 128). London: Department of Work and Pensions, 2004. (http://www.dwp.gov.uk/asd/asd5/index).
24 **De Boer AGEM**, van Lanschot JJB, Stalmeier PFM, *et al.* Is a single-item visual analogue scale as valid, reliable and responsive as multi-item scales in measuring quality of life? *Quality of Life Research* 2004;**13**:311–20.
25 **Fayers PM**, Hand DJ. Causal variables, indicator variables and measurement scales: an example from quality of life. *Journal of the Royal Statistical Society* 2002;**165**:1–21.
26 **Manning WG**, Newhouse JP, Ware JE. The status of health in demand estimation: beyond excellent, good, fair and poor. In: Fuchs VR, ed. *Economic aspects of health.* Chicago, IL: University of Chicago Press, 1982.
27 **McHorney CA**, Ware JE, Rogers W, *et al.* The validity and relative precision of MOS short- and long- form health status scales and Dartmouth COOP charts: results from the medical outcomes study. *Med Care* 1992;**30**:MS253–65.
28 **Ware JE**, Dewey JE. Health status and outcomes assessment tools. *International Electronic Journal of Health Education* 2000;**3**(special):138–48.
29 **Streiner DL**, Norman GR. In: *Health measurement scales: a practical guide to their development and use.* 3rd ed. Oxford: Oxford University Press, 2003.
30 **Nunnaly J**, Bernstein I. *Psychometric theory.* 3rd ed. New York: McGraw Hill, 1994.
31 **Rust J**, Golombok S. *Modern psychometrics. The science of psychological assessment.* 2nd ed. London: Routledge, 1999.

Public health interventions

Efficacy, effectiveness, and the evaluation of public health interventions

Mauricio L Barreto

Epidemiologists face a permanent challenge towards improving the design, analysis, interpretation, and reporting of observational and evaluative studies.

Based on the analysis of a sample of articles that reported the results obtained in observational epidemiological studies published in major journals, Pocock *et al*[1] raised *''serious concerns regarding inadequacies in the analysis and reporting of epidemiological publications''*, thereby intensifying the outcry regarding the low quality of these papers and the need for guidelines regulating their publication.[2] Moreover, epidemiological studies that fail to identify important associations or in which associations that were reported are later shown in effect not to exist abound in the literature. These errors often result in serious consequences, for example, when they involve the evaluation of interventions that cause adverse effects on human health. One recent case refers to the association between hormone replacement therapy (HRT) and cardiovascular disease (CVD), in which various observational studies systematically pointed in one single and erroneous direction—to the protective role of HRT in the occurrence of CVD.[3] Surprisingly, two recently published randomised controlled trials (RCT) showed completely contrasting results—a harmful effect of HRT on the occurrence of CVD.[4 5] There is no doubt that this ''mistake'' has had serious implications in women's health as for several years millions of women worldwide were prescribed with HRT without doctors and patients being aware of the harm it could cause.

As epidemiological knowledge is predominantly built on results obtained from observational studies, the implications of wrong findings originating from these studies have been constraining both for epidemiologists and for those who make use of epidemiological knowledge. Consequently, demands are quite rightly being made for reconsideration of issues concerning the relation between the outcomes of experimental and observational epidemiological studies. Discussions have covered conceptual and methodological aspects and more pragmatically the implications of the two different approaches in medicine and public health.

With respect to the evaluation of medical technologies (medicines, vaccines, etc), there is a general consensus that RCTs are the gold reference standard; however, this consensus is commonly extrapolated to the idea that, just as medical interventions, public health ones not submitted to randomised trials are unworthy of consideration as such, and it is recommended: *''to reject the scientific double standard of what constitutes acceptable evidence of efficacy for clinical versus public health interventions''*.[6]

This question, despite all its importance for public health practices, has received sparse attention from epidemiologists and public health practitioners. While the modern practice of medicine is centred on interventions developed as a result of biomedical research, the same does not occur in public health. The central aim of any public health intervention must be to modify the health of populations, eventually by reducing the harmful effects of morbid occurrences but principally by reducing the occurrence rates of these events—that is, their incidences. With respect to the goal of reducing incidence, biomedicine so far has made available to public health a set of interventions that, although relevant, are limited to only a few of the problems tormenting the health of the human populations. Vaccines are perhaps the most important biomedical technology offered to the public health intervention arsenal in present times; however, they are restricted to the infectious diseases and to just a few of them. The great majority of potential public health interventions whether behavioural, environmental, or social that could have a modifying effect on the population health in terms of a reduction in the incidence of specific or unspecific morbid events are outside the sphere of biomedicine.

Another no less important aspect is that public health interventions, even if applicable in individuals, need to be applied over populations to be effective. To achieve their aims they have to be organised into programmes running in the frame of established health policies and at the same time to have several other characteristics apart from efficacy.