Quality of Life Research, 4, pp. 293–307

Individual-patient monitoring in clinical practice: are available health status surveys adequate?

C. A. McHorney* and A. R. Tarlov

Departments of Preventive Medicine and Medicine, University of Wisconsin-Madison Medical School and the Health Services Research and Development Program, William S. Middleton Memorial Veterans Hospital, Madison, WI (C. A. McHorney); The Health Institute of the New England Medical Center, The Harvard School of Public Health and Tufts University, Boston, MA (A. R. Tarlov)

Interest has increased in recent years in incorporating health status measures into clinical practice for use at the individual-patient level. We propose six measurement standards for individual-patient applications: (1) practical features, (2) breadth of health measured, (3) depth of health measured, (4) precision for cross-sectional assessment, (5) precision for longitudinal monitoring and (6) validity. We evaluate five health status surveys (Functional Status Questionnaire, Dartmouth COOP Poster Charts, Nottingham Health Profile, Duke Health Profile, and SF-36 Health Survey) that have been proposed for use in clinical practice. We conducted an analytical literature review to evaluate the six measurement standards for individual-patient applications across the five surveys. The most problematic feature of the five surveys was their lack of precision for individual-patient applications. Across all scales, reliability standards for individual assessment and monitoring were not satisfied, and the 95% CIs were very wide. There was little evidence of the validity of the five surveys for screening, diagnosing, or monitoring individual patients. The health status surveys examined in this paper may not be suitable for monitoring the health and treatment status of individual patients. Clinical usefulness of existing measures might be demonstrated as clinical experience is broadened. At this time, however, it seems that new instruments, or adaptation of existing measures and scaling methods, are needed for individual-patient assessment and monitoring.

Key words: Clinical practice; health status assessment; individual patient applications; measurement standards reliability; score distributions; validity.

Preparation of this paper was supported by a grant from the Functional Outcomes Program of the Henry J. Kaiser Foundation at the Health Institute, New England Medical Center, Boston, MA (Grant Number 93-002), and by the Department of Veterans Affairs.

* To whom correspondence should be addressed.

Introduction

In the last 25 years, general health status measures have been used in a wide range of group-level applications including: (1) descriptions of health profiles for patients differing in diagnosis,[1,2] disease severity[3,4] and treatment regimen;[5,6] (2) evaluations of the relative benefits of different treatments;[7,8] (3) comparisons of health outcomes across different health care delivery systems;[9,10] (4) assessments of health policy initiatives[11–13] and (5) measurement of general population health.[14–16] Diverse health status measures (such as the Quality of Well-Being Scale,[17] the Sickness Impact Profile,[18] the McMaster Health Index Questionnaire,[19] the RAND Health Insurance Experiment surveys,[20] the Nottingham Health Profile,[21] the Duke Health Profiles[22,23] and surveys from the Medical Outcomes Study[24–26]) have been designed for and used in group-level applications.

Recently, however, managed care plans, quality improvement teams, and practising clinicians have displayed growing interest in using health status measures in clinical practice for individual-patient assessment and treatment monitoring,[27–29] where the unit of analysis is the individual patient rather than an aggregation of patients. Incorporation of general health measures in clinical practice theoretically could serve numerous purposes:[30–32] (1) describing patients' overall state; (2) screening for incipient disease; (3) assessing needs; (4) setting treatment goals; (5) monitoring disease progression; (6) monitoring response to treatment; (7) improving physician–patient communication and (8) standardizing interactions between health care providers and patients.

Why is there growing interest in using health status measures for individual-patient applications in

clinical practice? First, the modern medical paradigm, which focuses heavily on anatomical and pathophysiological parameters, is increasingly being recognized as an incomplete model for understanding the human experience of chronic disease. Indeed, in response to the seeming narrowness of the current medical paradigm, there has been increased call in the last ten years for a biopsychosocial model of medicine.[33–35] Second, research conducted over the last decade has consistently demonstrated poor correspondence between physician and patient ratings of functional status, emotional well-being and quality of life (QOL).[36–40] Health care providers tend to underestimate patients' functional disabilities, particularly psychosocial disabilities. Third, as the population continues to age and chronic disease increasingly threatens individual and population health, QOL concerns become paramount for the individual patient, the family and society at large. Recognizing the challenge that chronic disease represents, leading clinical spokespersons have advocated the maintenance and enhancement of patient function and well-being as an essential goal of medical care.[41–44]

Numerous position papers have proposed measurement standards to evaluate and select health status measures for group-level applications.[45–47] In this paper we propose measurement standards for individual-patient applications, and we evaluate five general health status surveys proposed for use in clinical practice. Our principal goal is to stimulate critical appraisal of the applicability of existing measures to individual patients and to encourage new measurement development, or adaptation of existing measures and scaling methods, to meet the growing demand for routine functional assessment in clinical practice.

Methods

Health surveys compared

We reviewed the literature on general health status measures and identified five general health surveys proposed for individual-patient applications. Two surveys—the Functional Status Questionnaire (FSQ)[48] and the Dartmouth COOP Poster Charts (COOP)[49]—were designed for individual-patient use in routine clinical practice. The FSQ is a 34-item questionnaire designed to screen for disability and to monitor change in functioning in primary care. The COOP charts were constructed for screening purposes; other applications suggested by its developers include assessment and monitoring of function, diagnosing disease and planning care. Each of the nine COOP charts enumerates five scale levels, using both written words and an illustration.

Three surveys—the Nottingham Health Profile (NHP),[21] the Duke Health Profile (DUKE),[23] and the SF-36 Health Survey (SF-36)[26]—were designed for group-level applications but have also been recommended for use in clinical practice. The NHP contains 38 items and was devised for use in health services planning and evaluation, population surveys, health promotion and clinical practice.[50] The DUKE is a 17-item questionnaire constructed for use in research, health promotion and clinical practice. The SF-36 is a 36-item survey designed for use in health policy evaluations, clinical research and clinical practice.

Measurement standards for individual-patient applications

If health status measures are to be used at the individual-patient level, they should meet several essential measurement standards. First, they should be brief and be easy to administer, score and interpret[30,31] because most clinical encounters are brief (9–17 minutes on average[51]) and because functional assessment is currently not a reimbursable expense. Second, to be useful for individuals differing in age, diagnosis, severity and comorbidity, measures should tap a variety of health concepts, each of which should assess the full range of health, from disability to well-being. Third, to yield clinically-useful descriptions of function across diverse patient groups, both at a single point in time and over time, measures should exhibit minimal floor and ceiling effects (percentage of the sample achieving the worst and best possible scores, respectively). Fourth, measures should yield highly accurate and precise scores that have a small standard error of measurement for use in clinical decision-making.[30,52] Fifth, measures should be highly reproducible over time and have a small standard error of measurement for use in longitudinal monitoring.[30,31] Finally, measures should be valid indicators of the constructs they are hypothesized to represent, show sensitivity to clinical change and have evidence of validity for individual-patient applications.[53,54]

Methods of analysis

We conducted an analytical literature review to evaluate these six measurement standards for individual-patient applications across the five surveys. To illustrate some measurement standards, we also used data

on the COOP charts collected for the Medical Outcomes Study (MOS). Evaluation criteria for each measurement standard are described below.

Practical features. We compared the number of items and survey administration time to evaluate criteria pertaining to brevity and ease of administration. Surveys with fewer items and shorter administration times may be more practical for routine use in clinical practice. However, as discussed later, trade-offs exist between survey length, breadth and depth of measurement, distributional characteristics and measurement precision. Surveys that take 5–15 minutes to complete by self-administration are defined as meeting practical standards. To assess ease of scoring, we compared scaling methods. Measures that require computers to generate scores may be less practical for routine use in clinical practice, although the widespread availability of computer technology makes this criterion less burdensome today than even a few years ago.

Breadth of health measured. To assess the breadth of health measured, we compared the content of the five health surveys. Based on a review of definitional standards,[55,56] surveys were defined as meeting content standards if they included scales tapping physical, role and social functioning and mental health.

Depth of health measured. To assess the depth of health measured, we compared the prevalence of floor and ceiling effects. Surveys with no floor or ceiling effects were defined as exceeding standards pertaining to score distributions, those with small floor or ceiling effects (1%–15%) met standards, and surveys with moderate floor or ceiling effects ($> 15\%$) failed to meet standards.

Measurement precision (cross-sectional). We compared reliability estimates and their associated standard errors of measurement to assess measurement precision at a single point in time. For all surveys except the COOP charts, we reported internal-consistency reliability. For the COOP charts, we reported alternate-forms reliability, which is a lower-bound estimate of reliability.[25] Recommended reliability standards for individual-level applications range from a low of 0.90[57] to a high of 0.95, which is the desired standard.[57,58] Reliabilities falling within or exceeding this range satisfy recommended standards.

The standard error of measurement (SEM) is the standard deviation of an individual score and it is the most useful reliability estimate for individual-level applications.[59] The SEM reflects both reliability and variance, as defined by $\text{sd} \times \sqrt{1 - \text{reliability}}$. A perfectly reliable instrument has a SEM of zero, i.e. the observed score is the true score. As one departs from perfect reliability, the 95% confidence interval (95% CI) can be calculated to gauge the certainty with which an observed score can be viewed as measuring the 'true' score. The 95% CI is an index of the random variation expected if an individual were tested repeatedly with equivalent forms.[60]

Measurement precision (longitudinal-monitoring). For reproducibility, we compared test–retest reliability estimates (2–4 weeks) and their associated SEM. Test–retest reliabilities greater than 0.90 were defined as satisfying recommended standards.

We used the 95% CI of the SEM (based on test–retest reliability) to gauge the likelihood that a difference between two scores of an individual is attributable to random error rather than true change. If the difference between two scores for an individual lies within the 95% CI, we can be reasonably certain that the observed score change is random error. Use of the SEM to define clinically-meaningful change is also known as the Reliable Change index (RC).[61] A more conservative test, based on the standard error of the difference between two scores, is known as the Significant Change index (where $\text{SC} = \sqrt{2}(\text{RC})$).[62]

Validity. There are neither clear-cut nor quantitative standards for validity as there are for reliability. However, measures used for individual-patient applications should at least exhibit satisfactory convergent and discriminant validity, tests of which gauge measurement integrity and provide useful information about score interpretation. Measures should also have evidence pertaining to responsiveness or sensitivity to longitudinal change. Finally, because validity, like reliability, is application and population specific rather than an inherent attribute of a measure, evidence of validity as it pertains to individual-patient applications (e.g. screening and decision-making[57]) should be available.

Results

Practical features

As Table 1 shows, the five surveys differ somewhat in practical features. The COOP charts have the fewest number of items ($k = 9$), followed by the DUKE ($k = 17$), with the other three surveys having 34–38 items. Consistent with the number of items, average self-administration time varies across surveys, with the COOP charts and the DUKE having the

Table 1. Comparison of practical features across surveys

Feature	NHP	FSQ	COOP	DUKE	SF-36
No. items	38	34	9	17	36
Average self-administration time (min)	10	15	6	5	10
Scaling method	Thurstone	Likert for 6; scaling not needed for 6	Not needed	Likert	Likert

NHP = Nottingham Health Profile
FSQ = Functional Status Questionnaire
COOP = Dartmouth COOP Charts
DUKE = Duke Health Profile
SF-36 = MOS SF-36 Health Survey

Table 2. Comparison of content across surveys

Health concept	NHP	FSQ	COOP	DUKE	SF-36
Physical functioning	×	×	×	×	×
Social functioning	×	×	×	×	×
Mental health	×	×	×	×	×
Pain	×		×	×	×
Health perceptions		×	×	×	×
Role functioning		×	×		×
Vitality	×				×
Disability		×		×	
Change in health			×		×
Sleep	×				
Sexual functioning		×			
Depression				×	
Anxiety				×	
Self-esteem				×	
Quality of interaction		×			
Social support			×		
Overall life quality			×		

shortest administration time (5–6 minutes),[63,64] followed by the SF-36 and the NHP (10 minutes each),[26,65] and the FSQ (15 minutes).[48] Scaling techniques are not required for the COOP charts because they are single-item measures. The other four surveys require hand scoring or computerized aggregation of items into scale scores.

Breadth of health measured

The five surveys differ somewhat in the number of discrete health scales enumerated (Table 2): the NHP yields six scales, the SF-36 eight, the COOP charts nine, the DUKE nine and the FSQ twelve. (Table 2 lists eight of the twelve FSQ scales; there are two scales each for disability, physical, social and role functioning. Consistent with the World Health Organization's definition of health,[66] all five surveys assess physical functioning, social functioning and mental health. Three surveys measure role functioning. All but the FSQ evaluate bodily pain, and all but the NHP appraise health perceptions. However, each survey assesses somewhat different secondary health concepts. For example, the FSQ taps sexual functioning, the NHP assesses sleep, the COOP charts tap social support and the DUKE includes self-esteem.

Depth of health measured

Table 3 summarizes the prevalence of floor and ceiling effects across the five surveys. Data are the range of floor and ceiling effects across all scales within a

Table 3. Comparison of floor and ceiling effects across surveys

	NHP	FSQ	COOP	DUKE	SF-36
% Floor	0–11	—*	1–6	0–19	0–24
% Ceiling	48–78	—*	12–66	1–72	1–56

* Data not published

given survey; Appendix Tables A.1–A.4 (see Appendix) provide floor and ceiling effects for each scale within each survey (where available). As Table 3 shows, most surveys exhibit minor floor effects. The NHP assesses negative dimensions of health. As a result, floor effects were largely absent in a general population (only one scale, energy, had floor effects exceeding 1%).[67] Floor effects were small for the COOP charts in the MOS (a study of chronically ill patients), ranging from 1–6% with an average of 2.5% across all nine charts. In a sample of primary care patients, floor effects for the DUKE ranged from 0–19% and averaged 2.9% across all 10 scales (only the pain scale had floor effects exceeding 5%).[68] In the MOS, floor effects for the SF-36 were very small ($\leqslant 1\%$) except for the two role disability scales (24% for role disability–physical and 18% for role disability–emotional).[69] Floor effects have not been published for the FSQ.

Because all five surveys tend to represent health as the absence of limitations, it is not surprising that ceiling effects are more prevalent and problematic than are floor effects. Substantial ceiling effects have been reported for the NHP, ranging from 48–78% in community-based studies.[50,67,70] Ceiling effects for the NHP in a patient-based sample have not been reported. Ceiling effects for the COOP charts in the MOS ranged from 12% (change in health) to 66% (social functioning) and averaged 37% across the nine charts. Ceiling effects for the DUKE in a patient sample ranged from 1% (general health) to 72% (disability) and averaged 20% across the ten scales.[68] A range of ceiling effects from 1% (vitality) to 56% (role disability–emotional) were reported for the SF-36 among MOS patients, with an average of 23% across all eight scales.[69] Ceiling effects have not been published for the FSQ.

Measurement precision (cross-sectional). The top panel of Table 4 summarizes estimates of precision (reliability and the 95% CI of the SEM) for cross-sectional applications. Ranges of reliability estimates across all scales within a given survey are presented; reliability estimates for specific scales within each survey (where available) are reported in Tables A.5–A.9. Few surveys meet minimum reliability standards for individual-level applications. Only two SF-36 scales (physical functioning and mental health) meet the lower-bound standard of 0.90, and none of the scales meet the recommended standard of 0.95. In short, none of the surveys achieve the degree of reliability that would be desirable for individual assessment and decision-making.

Failure to meet reliability standards is reflected in the 95% CIs of the SEM. As summarized in Table 4 and detailed in Tables A.5–A.9, the 95% CIs are wide, often comprising up to one-third or more of the score distribution. For example, the 95% CI of an obtained physical functioning score is ± 45 points for the COOP chart, ± 32 points for the NHP, ± 25 points for the DUKE, and ± 14 points for the SF-36. Because the lower and upper limits of the 'true' physical functioning score are so wide, none of these scales may be capable of providing an accurate reading of physical functioning for most individual-patient, clinical decision-making purposes.

Measurement precision (longitudinal-monitoring). The bottom panel of Table 4 presents precision estimates for longitudinal monitoring. These are ranges of test–retest reliability estimates and the associated 95% CI of the SEM; precision estimates for specific scales within each survey (where available) are reported in Tables A.5–A.9. Again, the reliabilities of all scales fall far below the 0.90 to 0.95 standard, and the 95% CIs are very wide. Using physical functioning as an example, an individual patient's score would have to change by at least 32 points on the COOP chart, 23 points on the SF-36, 22 points on the DUKE, and 16 points on the NHP for the change to be considered statistically significant. In those infrequent clinical situations where only large changes in functioning and well-being need be detected, perhaps these wide confidence intervals would be acceptable.

Validity. Years of accumulated research have provided evidence pertaining to the validity of each of the

Table 4. Comparison of reliability estimates across surveys

	NHP	FSQ	COOP	DUKE	SF-36
Cross-sectional precision†	0.34–0.81	0.64–0.82	0.24–0.54	0.55–0.78	0.78–0.93
95% CI of the SEM	23–41	17–41	36–46	14–25	13–32
Longitudinal precision‡	0.77–0.85	*	0.42–0.88	0.30–0.78	0.60–0.81
95% CI of the SEM	16–30	*	20–50	14–50	19–47

* Data not published
† Internal-consistency reliability except for COOP, which is alternate-form reliability
‡ Test–retest reliability (2 to 4 weeks)

surveys. For example, NHP scales correspond well with other health indicators[70–76] and discriminate between well and sick populations.[77–81] Moderate correlations between the FSQ scales and other health constructs have been reported, thus providing evidence of construct validity.[48,82–84] The COOP charts exhibit good convergent and discriminant validity,[49,63,85] show moderate correlations with other health constructs[63,85,86] and display some evidence of predictive validity.[87] The convergent and discriminant validity of the DUKE is suggested by correlations with other health measures[23,88] and its scales discriminate between clinical groups.[23,89,90] The predictive validity of the DUKE *vis a vis* future utilization, costs and illness severity has also been documented.[90,91] The SF-36 scales correspond well with other health measures,[70,92–96] discriminate between well and sick populations[97] and are sensitive to differences in disease severity.[40,98–100] Not all of the surveys evaluated have been assessed for responsiveness to change and for those that have, mixed results have been observed.[83,84,97,101–110]

Summary of results. Table 5 summarizes the six broad measurement criteria for individual-patient applications (practical features, breadth and depth of health measured, measurement precision for cross-sectional and longitudinal monitoring, and validity), our proposed standards and an overall evaluation for each survey on each standard (A = excellent—standard exceeded; B = adequate—standard minimally met for most or all scales in a survey; C = inadequate—standard insufficiently met for most or all scales in a survey; D = very poor—standard not met for any scale in a survey).

The COOP charts excelled in all standards pertaining to practical features: they were the shortest in length, the briefest to complete and the easiest to score and interpret. The FSQ also ranked high in these criteria because six of its twelve scales are single-item measures while the FSQ, the NHP and the SF-36 did not differ from one another.

In terms of the breadth of health concepts measured, all five surveys assessed physical functioning, social functioning and mental health, but only three measured role functioning as well. Aside from these concepts, there was considerable variability in selected secondary concepts. In terms of depth of measurement, floor effects were far less prevalent and problematic than were ceiling effects.

The most problematic feature of the five surveys was their lack of precision for individual-patient applications. Across all scales, reliability standards for individual applications were not satisfied. Moreover, the 95% CIs were wide, thus rendering tentative (at best) clinical conclusions about an individual's observed score at a point in time or changes in an individual's score over time.

The process of validity assessment involves validating the use of a measure, not the measure itself.[53,111] Although each of the five surveys has accumulated considerable evidence of validity for group-level applications, there was little evidence of their validity for screening, diagnosing disease or monitoring individual patients. Of the five surveys evaluated, only the mental health scale included in both the SF-36 and FSQ (the MHI-5) and the mental health scales included in the DUKE had evidence of validity directly pertaining to their use as clinical tools for screening or diagnosis.[90,112] Evidence pertaining to responsiveness or longitudinal assessment, was mixed across each of the surveys and mixed across scales within surveys.

Discussion

A proliferation of standardized measures based on patient self-report that assess functional status, emotional well-being and subjective perceptions of health has occurred in the last 20 years. These measures have been used in a wide range of group-level applications. However, interest has increased in incorporating health status measures into clinical practice for routine use at the individual-patient level.[29,49,113,114] Growing

Table 5. Summary of measurement standards and survey instrument performance

Measurement criterion	Proposed standard	Survey Instrument Grade				
		NHP	FSQ	COOP	DUKE	SF-36
Practicality	5–15 minutes	A	A	A	A	A
	Ease of scoring	B	A	A	B	B
Breadth of health measured	Physical, role, social, mental	B	A	A	B	A
Floor effects	< 15%	A	?	A	B	B
Ceiling effects	< 15%	D	?	C	C	C
Precision (cross-sectional)	0.90–0.95 reliability	D	D	D	D	C
Precision (longitudinal)	0.90–0.95 reliability	D	?	D	D	D
Validity	Group level	A	A	A	A	A
	Individual level	D	C	C	C	C

Performance standards. A = excellent—standard exceeded; B = adequate—standard minimally met across most or all scales in a survey; C = inadequate—standard insufficiently met for most or all scales in a survey; D = poor—standard not met for any scale in a survey; ? = data not published

clinical enthusiasm for routine functional assessment,[115] yet disappointing evidence of its utility in clinical practice,[116–119] prompted us to propose measurement standards for individual-patient applications.

The health status surveys examined in this paper may not be suitable for monitoring the health and treatment status of most individual patients, particularly those with chronic disease. Their shortcomings for individual-patient monitoring have been suggested in this paper relative to reliability, the range of health states assessed and validity. The surveys do not meet minimum standards of reliability for individual-patient assessment and monitoring. The confidence intervals for most scales are so wide as to render the instruments insufficiently sensitive for detecting levels of functional disability and functional change that would be diagnostically and therapeutically useful in medical practice.

With few exceptions, current surveys do not measure the full range of health, thus yielding noteworthy ceiling effects. The consequences of significant ceiling effects for individual-patient assessment are two-fold: (1) measurement of improvement from the baseline perfect score is not possible; and (2) false-negative case-finding outcomes are likely when a patient is deemed to have perfect functioning on a measure that only assesses severe dysfunction.

It is essential that measures used in individual-patient applications have evidence of validity for cross-sectional case-finding and longitudinal monitoring. In terms of longitudinal assessment, the unit of analysis in group-level tests of responsiveness is typically the mean difference in scores between two (or more) assessments. Although a group may improve on average as a result of treatment, that average is composed of individual variation in deterioration, stability and improvement. For example, in a recent study of elective percutaneous coronary revascularization, the average improvement in the intervention group in physical functioning, as measured by the SF-36, was 19.1 points, a statistically significant finding.[120] However, when the 95% CI of the SEM was used to define individual-level change, 7% of individuals experienced significant decline in scores, 45% remained stable and 48% improved. Thus, evidence of a measure's responsiveness to change for a group does not necessarily have commensurate interpretation for individual patients.

In terms of validity for cross-sectional case-finding, only the MHI-5 and the DUKE mental health scales have clinically-established cut-points or thresholds that advise the clinician of functional perturbations worthy of further attention. Representative norms can facilitate clinical interpretation of scores for individual patients. However, because the 95% CIs can be wide (e.g. range = 13–32 points and average 21 points across all eight SF-36 scales for general population norms),[16] it is likely that clinically and socially relevant differences in health would exist between an individual patient and the norm that would be deemed statistically equivalent. Of course, one could use less stringent criteria and take a greater chance of making an inaccurate comparison. In lieu of mean scores obtained from norms, reference values could be used, such as percentile scores, to facilitate clinical interpretation of individual scores for case-finding and monitoring purposes.[121]

The appropriate use of norms and percentiles for clinical interpretation of individual-patient scores rests upon three essential attributes of the normative group: (1) its size; (2) its representativeness and (3) its comparability to the individual patient.[122] The first two requirements are fairly self-evident: norms based on a

small or unrepresentative sample may be inaccurate or, worse, biased if there was non-random selection of 'good' or 'bad' normative subjects. The third requirement is less intuitive. Norms are only meaningful to the extent that the individual whose score is being interpreted belongs to the normative population.[123] If the norm and the individual patient differ appreciably in sociodemographic characteristics or clinical case mix, norms can be of limited or even misleading value because one or both samples are biased in characteristics unrelated to the underlying construct of interest. Thus, unless the norm and individual patient are roughly comparable, one is comparing apples to oranges.

Many trade-offs exist in the quest for both measurement simplicity and precision for individual-patient applications. On one hand, score distributions, reliability and the precision of an individual score favour longer measures.[92] On the other hand, ease of administration, low respondent burden, low costs of data collection and scoring and ease of score interpretation favour the briefest of measures. Like others,[30,124] we maintain that measurement precision should be the first priority in individual-patient applications. Accordingly, measures used for individual assessment must achieve high levels of reliability.

Improvements in reliability could be accomplished in several ways including: (1) using more clearly written items with universally-understood words; (2) selecting items that are normally distributed, have adequate variance and have similar difficulty; (3) using categorical rating scales instead of dichotomous response choices; (4) increasing test length and (5) measuring unidimensional (homogeneous) rather than multidimensional (heterogeneous) constructs. Each of these standard elements of test construction will, to one degree or another, increase internal-consistency reliability.

Another way to improve reliability is to aggregate individual composite-scale scores into a summary scale. Aggregate indices are not new in health status assessment. For example, the Quality of Well-Being Scale[17] combines dimensions of functional status with symptom complexes, the Sickness Impact Profile[18] yields a summary physical score, a summary psychosocial score and an overall scale that combines physical and psychosocial health, and the Duke Health Profile[23] averages its physical, social and mental health scales into an overall general health scale.

Recent experiences in aggregating composite-scale scores derived from MOS surveys into summary indices[125–127] have yielded reliability estimates that meet minimum standards for individual-patient applications (e.g. internal-consistency reliability estimates of 0.94 for the Perceived Health Index[126] and 0.92 and 0.91 for the MOS Physical and Mental Summary Measures,[127] respectively). Although summary measures are an important step toward achieving desired levels of internal-consistency reliability and reducing floor and ceiling effects, they are by definition associated with a loss of potentially clinically-relevant information about different aspects of a given patient's functioning and well-being.

Even with improved precision, advances in scaling methods may be warranted to improve the clinical interpretation and thus practical clinical utility of individual-patient scores. Likert and Thurstone scaling methods yield scores that are not easily interpreted. Scores between the lowest and highest possible values cannot be aetiologically interpreted because a specific score can be achieved by myriad combinations of item responses. For example, there are 2850 possible ways to obtain a score of 70 on the SF-36 physical functioning scale, although not all combinations are observed in a given sample. Scores obtained from summary indices are also uninterpretable as to cause because a given score can be achieved by innumerable combinations of weighted component-scale scores. Further, interpreting the meaning of changes in summary scores is not feasible because a change of a given unit (e.g. 5 points) can be achieved by countless combinations of worsening, stability or improvement on each component-scale score. In short, the inability to easily understand the pattern of item responses (i.e. the specific disabilities) or composite-scale scores that yields a given individual's score compromises the clinical utility of functional assessment in clinical practice. However, as in all other tests and procedures used in clinical practice, a given test result must be interpreted in conjunction with a clinician's findings for best results.

Use of an alternative scaling technique, item response theory,[128] holds promise for improving the interpretability of individual scores. Rasch scaling yields a score based on an ordered continuum (hierarchy) of items. An advantage of Rasch scaling is that the item combinations that yield a given score are determinable, as are individual-patient deviations from the empirically-based expected item hierarchy.[129] One disadvantage of Rasch scaling is that items must satisfy several difficult criteria, including a hierarchial structure and unidimensionality.[130] Recent experiences using Rasch scaling with the MOS physical functioning scale have yielded useful information about differential item difficulties (gaps in the physical functioning continuum) and item redundancies that may prove useful in future work.[130]

It is questionable that generic measures of

individual-level health status alone will provide all the information needed for effective clinical decision-making. Disease-specific measures that focus on symptom status and specific limitations (e.g. range of motion of the hip joint) may be required, but their psychometric properties will also have to meet a high standard if used for individual-patient assessment and monitoring. Disease-specific measures that explicitly incorporate individually-weighted preferences for health states[131–133] could also prove useful in clinical practice, particularly in assessing individual needs and setting treatment goals that are compatible with individual health-state preferences.

Our comments thus far address problems with precision and scaling but attempts to remedy the shortcomings could aversely affect considerations for routine use in everyday clinical practice. To reduce concerns about office burden, patients could complete health status measures at home and transmit them to the office electronically or by mail or telephone for scoring, interpretation and filing. The growing movement toward computerized-adaptive testing in the educational testing enterprise also holds great promise for individual-patient assessment since gains in efficiency can be achieved without a loss in precision. Development of computerized-adaptive testing for functional health assessment, however, would require a large computerized bank of questionnaire items, ability-matched item administration at the individual-respondent level and alternatives to traditional scaling approaches (e.g. use of Item Response Theory).[134]

The field of health status assessment has successfully developed many new tools, both generic and disease-specific, for group-level applications—particularly clinical research and trials. The mission of this paper was to evaluate the applicability of five of these survey tools to individual-patient assessment and monitoring in clinical practice. We determined that these surveys did not sufficiently meet our proposed measurement standards for individual-patient assessment. But are our proposed standards too high? Clearly some commonly-used clinical tests[135,136] fall short of minimum standards of reliability[57,58,137] for individual-patient assessment. Similarly, some clinical tests exhibit noteworthy floor and ceiling effects while others differ greatly in their sensitivity, specificity and false-positive rates.[138,139] However, clinicians do appear to develop an intuitive understanding of clinical test results that is shaped by years of formal graduate medical education and clinical experience in interpreting scores in everyday patient care.

Perhaps applications of health status measures in clinical practice have been too few to date to sufficiently evaluate their potential clinical utility. Clearly, routine use of health status measures in everyday clinical practice on an individual-patient basis has been a very recent phenomenon. Therefore, some might consider it premature to fault current measures on theoretical and statistical grounds when they may yield useful information for some individual clinicians and for some individual patients. Perhaps, as with common clinical tests, individual clinicians will differ greatly in the amount of measurement error or diagnostic uncertainty they are willing to accept in test results. Some clinicians may tolerate a considerable amount of error overall if a health status measure yields an occasional discovery of poor mental health or occasionally renders a prompt for more detailed assessment. However, because of the wide confidence intervals around an individual-patient score, it is unlikely that current measures will be useful for most longitudinal monitoring purposes, such as adjusting treatment regimens based on individual-patient health status scores in patients having diabetes, hypertension and other chronic diseases.

The goal of this paper was to appraise the suitability of currently-available health status measures for individual-patient assessment. As the analyses revealed, the tools available today are far from perfect for individual-patient assessment and particularly so for longitudinal individual-patient monitoring. In the absence of a 'perfect' tool for individual-patient assessment and monitoring, we offer three concluding recommendations. First, for those clinicians who are incorporating health status measures in their practice on an individual-patient basis, caution should be exercised in the interpretation of imprecise scores. Second, clinicians who use health status measures in their practice on an *individual-patient level* should report their experiences and findings to help move the field forward. Current users are in the best position to inform the field of gaps in content or problems in reliability, validity or responsiveness that can be subsequently redressed. Third, for instrument developers, the time is right to develop new measurement systems or to adapt the existing measures and scaling methods used for individual-patient assessment and monitoring. This will be necessary to satisfy with suitable precision and validity the growing demand for routine functional assessment in clinical practice.

Acknowledgements

The authors gratefully acknowledge comments provided on earlier drafts of the manuscript by Sonja Hunt, Ph.D., Eugene C. Nelson, DSc, George R. Parker-

son, M.D., M.P.H., John Ware, Ph.D., and William Rogers, Ph.D. and gratefully acknowledge the editorial assistance provided by Jennifer Lin and Cindi Birch. The views expressed in this manuscript are entirely those of the authors and are not intended to reflect the opinions of the commentators in any way.

References

1. Jenkinson C, Fitzpatrick R, Argyle M. The Nottingham Health Profile: an analysis of its sensitivity in differentiating illness groups. *Soc Sci Med* 1988; **27**: 1411–1414.
2. Stewart AL, Greenfield S, Hays RD, *et al.* Functional status and well-being of patients with chronic conditions: results from the Medical Outcomes Study. *JAMA* 1989; **262**: 907–913.
3. Alonso J, Anto JM, Gonzalez M, *et al.* Measurement of general health status of non-oxygen-dependent chronic obstructive pulmonary disease patients. *Med Care* 1992; **30 (Suppl)**: MS125–MS135.
4. Deyo RA, Inui TS, Leininger J, Overman S. Physical and psychosocial function in rheumatoid arthritis. *Arch Intern Med* 1982; **142**: 879–882.
5. Heaton RK, Grant I, McSweeney J, Adams K, Petty T. Psychological effects of continuous and nocturnal oxygen therapy in hypoxemic chronic obstructive pulmonary disease. *Arch Intern Med* 1983; **143**: 1941–1947.
6. Hart LG, Evans RW. The functional status of ESRD patients as measured by the Sickness Impact Profile. *J Chron Dis* 1987; **40 (Suppl)**: 117S–130S.
7. Bombardier C, Ware J, Russell IJ, *et al.* Auranofin therapy and quality of life in patients with rheumatoid arthritis: results of a multicenter trial. *Amer J Med* 1986; **81**: 565–578.
8. Canadian Erythropoietin Study Group. Association between recombinant human erythropoietin and quality of life and exercise capacity of patients receiving haemodialysis. *BMJ* 1990; **300**: 573–578.
9. Ware JE, Brook RH, Rogers WH, *et al.* Comparison of health outcomes at a health maintenance organization with those of a fee-for-service care. *Lancet* 1986; 1017–1022.
10. Retchin SM, Clement DG, Rossiter LF, *et al.* How the elderly fare in HMOs: outcomes from the Medicare competition demonstration. *Health Serv Res* 1992; **27**: 651–669.
11. Patrick DL, Erickson P. *Health Status and Health Policy: Allocating Resources to Health Care.* New York: Oxford University Press, 1993.
12. Kaplan RM. *The Hippocratic Predicament: Affordability, Access, and Accountability in American Medicine.* San Diego: Academic Press, 1993.
13. Newhouse JP, Manning WG, Keeler EB, Sloss EM. Adjusting capitation rates using objective health measures and prior utilization. *Health Care Fin Rev* 1989; **10**: 41–54.
14. Hunt SM, McEwen J, McKenna SP. Perceived health: age and sex comparisons in a community. *J Epid Commun Health* 1984; **38**: 156–160.
15. Brorsson B, Ifver J, Hays RD. The Swedish Health-Related Quality of Life Survey (SWED-QUAL). *Quality Life Res* 1993; **2**: 33–45.
16. McHorney CA, Kosinski M, Ware JE. Comparisons of the costs and quality of norms for the SF-36 Survey collected by mail versus telephone interview: results from a national survey. *Med Care* 1994; **32**: 551–567.
17. Patrick DL, Bush JW, Chen MM. Toward an operational definition of health. *J Health Soc Behav* 1973; **14**: 6–21.
18. Bergner M, Bobbitt RA, Carter WB, Gilson BS. The Sickness Impact Profile: development and final revision of a health status measure. *Med Care* **19**(8): 787–805.
19. Chambers LW. The McMaster Health Index Questionnaire: an update. In: Walker SR, Rosser RM, eds. *Quality of Life: Assessment and Application.* Lancaster: MTP Press Limited, 1988: 113–131.
20. Brook RH, Ware JE, Davies-Avery A, *et al.* Overview of adult health status measures fielded in RAND's Health Insurance Study. *Med Care* **17(Suppl)**; 1979: 1–131.
21. Hunt SM, McEwen J. The development of a subjective health indicator. *Soc Health Illness* 1980; **2**: 231–246.
22. Parkerson GR, Gehlbach SH, Wagner EH, *et al.* The Duke-UNC Health Profile: an adult health status instrument for primary care. *Med Care* 1981; **19**: 806–828.
23. Parkerson GR, Broadhead WE, Tse CJ. The Duke Health Profile: A 17-item measure of health and dysfunction. *Med Care* 1990; **28**: 1056–1072.
24. Stewart AL, Hays RD, Ware JE. The MOS Short-Form General Health Survey: reliability and validity in a patient population. *Med Care* 1980; **26**: 724–735.
25. Stewart AL, Ware JE. *Measuring Functioning and Well-Being: The Medical Outcomes Study Approach.* Chapel Hill, NC: Duke University, 1992.
26. Ware JE, Sherbourne CD. The MOS 36-Item Short-Form Health Survey (SF-36): I. Conceptual framework and item selection. *Med Care* 1993; **30**: 473–483.
27. Greenfield S. What's the next step for outcomes assessment? *The Internest* 1990; 6–10.
28. Wolfe F, Pincus T. Standard self-report questionnaires in routine clinical and research practice—an opportunity for patients and rheumatologists. *J Rheumat* 1991; **18**: 643–646.
29. Meyer KB, Espindle DM, DeGiacomo JM, *et al.* Monitoring dialysis patients' health status. *Am J Kidney Dis* 1993; **9**: 267–279.
30. Kane RA, Kane RL. *Assessing the Elderly: A Practical Guide to Measurement.* Lexington MA: 1981.
31. Applegate WB, Blass JP, Williams TF. Instruments for the functional assessment of older patients. *NEJM* 1990; **322**: 1207–1214.
32. Greenfield S, Nelson EC. Recent developments and future issues in the use of health status assessment measures in clinical settings. *Med Care* 1992; **30(Suppl)**: MS23–MS41.
33. Engel GL. The need for a new medical model: a challenge for biomedicine. *Science* 1976; 129–136.
34. Lipkin M, Quill TE, Napodano RJ. The medical interview: a core curriculum for residents in internal medicine. *Ann Intern Med* 1984; **100**: 277–284.
35. Fretwell MD. Comprehensive functional assessment (CFA) in everyday practice. In: Hazzard WR, Andes R, et al. eds. *Principles of Geriatric Medicine and Gerontology.* New York: McGraw-Hill 1990: 218–223.
36. Jachuck SJ, Brierly H, Jachuch S, Wilcox PM. The

effect of hypotensive drugs on the quality of life. *J Royal College Gen Pract* 1982; **32**: 103–105.
37. Patrick DL, Peach H, Gregg I. Disablement and care: a comparison of patient views and general practitioner knowledge. *J Royal College Gen Pract* 1982; **32**: 429–434.
38. Nelson EC, Conger B, Douglass R, *et al.* Functional health status levels of primary care patients. *JAMA* 1983; **249**: 3331–3338.
39. Calkins DR, Rubenstein LV, Cleary PD, *et al.* Failure of physicians to recognize functional disability in ambulatory patients. *Ann Intern Med* 1991; **114**: 451–454.
40. Nerenz DR, Repasky DP, Whitehouse FW, Kahkonen DM. Ongoing assessment of health status in patients with diabetes. *Med Care* 1992; **30(Suppl)**: MS112–MS124.
41. Cluff LE. Chronic disease, function, and the quality of care. *J Chron Dis* 1981; **34**: 299–304.
42. McDermott W. Absence of indicators of the influence of its physicians on a society's health. *Am J Med* 1981; **70**: 833–843.
43. Tarlov AR. The increasing supply of physicians, the changing structure of the health services system, and the future practice of medicine. *NEJM* 1983; **308**: 1235–1244.
44. Tarlov AR. The coming influence of a social sciences perspective on medical education. *Acad Med* 1992; **67**: 724–731.
45. Ware JE. Methodological considerations in the selection of health status assessment procedures. In: Wenger NK, Mattson ME, Furberg CD, Elinson J, eds. *Assessment of Quality of Life in Clinical Trials of Cardiovascular Therapies.* New York: Le Jacq Publishing, 1984: 87–111.
46. Bergner M, Rothman ML. Health status measures: an overview and guide for selection. *Ann Rev Pub Health* 1987; **8**: 191–210.
47. Hays RD, Anderson R, Revicki D. Psychometric considerations in evaluating health-related quality of life measures. *Qual Life Res* 1993; **2**: 441–449.
48. Jette AM, Davies AR, Cleary PD, *et al.* The Functional Status Questionnaire: reliability and validity when used in primary care. *J Gen Intern Med* 1986; **1**: 143–149.
49. Nelson E, Wasson J, Kirk J, *et al.* Assessment of function in routine clinical practice: description of the COOP chart method and preliminary findings. *J Chron Dis* 1987; **40(Suppl)**: 55S–63S.
50. Hunt SM, McEwen J, McKenna SP. *Measuring Health Status.* Dover NH: Croom Helm, 1986.
51. Greenwald HP, Peterson ML, Garrison LP, *et al.* Interspeciality variation in office-based care. *Med Care* 1984; **22**: 14–29.
52. Williams JI, Naylor CD. How should health status measures be assessed? Cautionary notes on procrustean frameworks. *J Clin Epidem* 1992; **45**: 1347–1351.
53. American Psychological Association. *Standards for Educational and Psychological Testing.* Washington, DC: American Psychological Association, 1985.
54. Overall JE. Contradictions can never a paradox resolve. *Appl Psych Meas* 1989; **13**: 426–428.
55. Bergner M. Measurement of health status. *Med Care* 1985; **23**: 696–704.
56. Patrick DL, Erickson P. What constitutes quality of life? *Qual Life Cardiovasc Care* 1988; **4**: 103–127.
57. Nunnally JC. *Psychometric Theory.* New York: McGraw-Hill, 1978.
58. Kaplan RM, Saccuzzo DP. *Psychological Testing. Principles, Applications, and Issues.* Pacific Grove, CA: Brooks/Cole Publishing Company, 1989.
59. Anastasi A. *Psychological Testing.* New York: Macmillan, 1990.
60. Thorndike RM, Cunningham GK, Thorndike RL, Hagan EP. *Measurement and Evaluation in Psychology and Education.* New York: Macmillan, 1991.
61. Jacobson NS, Follette WC, Revenstorf D. Psychotherapy outcome research: methods for reporting variability and evaluating clinical significance. *Behav Therapy* 1984; **15**: 336–352.
62. Christensen L, Mendoza JL. A method of assessing change in a single subject: an alteration of the RC Index. *Behav Therapy* 1986; **17**: 305–308.
63. Siu Al, Hays RD, Ouslander JG, *et al.* Measuring functioning and health in the very old. *J Gerontol* 1993; **48**: M10–M14.
64. Parkerson GR, January 14, 1994. [Personal communication]
65. Hunt SJ. Nottingham Health Profile. In: Walker SR, Rosser RM, eds. *Quality of Life: Assessment and Application.* Lancaster: MTP Press Limited, 1988: 165–169.
66. World Health Organization. *Constitution of the World Health Organization. Basic Documents.* Geneva, Switzerland: World Health Organization, 1948.
67. Anderson J, Sullivan F, Usherwood TP. The Medical Outcomes Study Instrument (MOSI)—use of a new health status measure in Britain. *Family Practice* 1990; **7**: 205–218.
68. Parkerson GR. January 21, 1994. [Personal communication]
69. McHorney CA, Ware JE, Lu JFR, Sherbourne CD. The MOS 36-Item Short-Form Survey (SF-36). III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. *Med Care* 1994; **32**: 40–66.
70. Brazier JE, Harper R, Jones NMB, *et al.* Validating the SF-36 health survey questionnaire: new outcome measure for primary care. *BMJ* 1992; **305**: 160–164.
71. Harwood RH, Gompertz P, Ebrahim S. Handicap one year after a stroke: validity of a new scale. *J Neurol Neurosurg Psychiatry* 1994; **57**: 825–829.
72. Hunt SM, McKenna SP, McEwen J, *et al.* The Nottingham Health Profile: subjective health status and medical consultations. *Soc Sci Med* 1981; **15A**: 221–229.
73. Leavy R, Wilkin D. A comparison of two survey measures of health status. *Soc Sci Med* 1988; **27**: 269–275.
74. Baum FE, Cooke RD. Community-health needs assessment: use of the Nottingham Health Profile in an Australian study. *Med J Australia* 1989; **150**: 583–590.
75. Alonso J, Anto JM, Moreno C. Spanish version of the Nottingham Health Profile: translation and preliminary validity. *AJPH* 1990; **80**: 704–708.
76. Permanyer-Miralda G, Alonso J, Anto JP, *et al.* Comparison of perceived health status and conventional evaluation in stable patients with coronary artery disease. *J Clin Epid* 1991; **44**: 779–786.
77. Doll HA, Black NA, Flood AB, McPherson K. Patient-perceived health status before and up to 12 months after transurethral resection of the prostate for benign prostatic hypertrophy. *Brit J Urol* 1993; **71**: 297–305.

78. Hunt SM, McKenna SP, McEwen J, *et al.* A quantitative approach to perceived health status: a validation study. *J Epid Commun Health* 1980; **34**: 281–286.
79. Hunt SM, McEwen J, McKenna SP, *et al.* Subjective health of patients with peripheral vascular disease. *The Practitioner* 1982; **226**: 133–136.
80. Ebrahim S, Barer D, Nouri F. Use of the Nottingham Health Profile with patients after a stroke. *J Epid Commun Health* 1986; **40**: 166–169.
81. Wiklund I, Romanus B, Hunt SM. Self-assessed disability in patients with arthrosis of the hip joint. *Int Disabil Studies* 1988; **10**: 159–163.
82. Einarsson G, Grimby G. Disability and handicap in late poliomyelitis. *Scand J Rehab Med* 1990; **22**: 113–121.
83. Cleary PD, Greenfield S, McNeil BJ. Assessing quality of life after surgery. *Controlled Clinical Trials* 1991; **12**: 189S–203S.
84. Katz JN, Larson MG, Phillips CB, *et al.* Comparative measurement sensitivity of short and longer health status instruments. *Med Care* 1992; **30**: 917–925.
85. Nelson EC, Landgraf JM, Hays RD, *et al.* The functional status of patients: how can it be measured in physicians' offices? *Med Care* 1990; **28**: 1111–1126.
86. Meyboom-De Jong M, Smith RJA. Studies with the Dartmouth COOP charts in general practice: comparison with the Nottingham Health Profile and the General Health Questionnaire. In: WONCA Classification Committee, eds. *Functional Status in Primary Care.* New York: Springer-Verlag, 1990: 132–149.
87. Siu A, Reuben D, Ouslander J, Osterwell D. Using multidimensional health measures in older persons to identify risk of hospitalization and skilled nursing placement. *Qual Life Res* 1993; **2**: 253–261.
88. Parkerson GR, Broadhead WE, Tse CJ. Comparison of the Duke Health Profile and the MOS short-form in healthy young adults. *Med Care* 1991; **29**: 679–683.
89. Parkerson GR, Broadhead WE, Tse CJ. Quality of life and functional health of primary care patients. *J Clin Epid* 1992; **45**: 1303–1313.
90. Parkerson GR, Broadhead WE, Tse CJ. Anxiety and depressive symptom identification using the Duke Health Profile. *J Clin Epidem,* in press.
91. Parkerson GR, Broadhead WE, Tse CJ. Health status and severity of illness as predictors of outcomes in primary care. *Med Care* 1995; **33(1)**: 53–66.
92. 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(Suppl)**: MS253–MS265.
93. Weinberger M, Samsa GP, Hanlon JT, *et al.* An evaluation of a brief health status measure in elderly veterans. *JAGS* 1991; **39**: 691–694.
94. Brazier J, Jones N, Kind P. Testing the validity of the Euroqol and comparing it with the SF-36 health survey questionnaire. *Qual Life Res* 1993; **2**: 169–180.
95. Garratt AM, Ruta DA, Abdalla MI, *et al.* The SF 36 health survey questionnaire: an outcome measure suitable for routine use within the NHS? *BMJ* 1993; **306**: 1440–1444.
96. Jenkinson C, Coulter A, Wright L. Short form 36 (SF 36) health survey questionnaire: normative data for adults of working age. *BMJ* 1993; **306**: 1437–1440.
97. Phillips RC, Lansky DJ. Outcomes management in heart valve replacement surgery: early experience. *J Heart Valve Dis* 1992; **1**: 42–50.
98. Vickrey BG, Hays RD, Graber J, *et al.* A health-related quality of life instrument for patients evaluated for epilepsy surgery. *Med Care* 1992; **30**: 299–319.
99. Wells KB, Burnam MA, Rogers W, Camp P. The course of depression in adult outpatients: results from the Medical Outcomes Study. *Arch Gen Psychiatry* 1992; **49**: 788–794.
100. 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**: 247–263.
101. Caine N, Harrison SCW, Sharples LD, Wallwork J. Prospective study of quality of life before and after coronary artery bypass grafting. *BMJ* 1991; **302**: 511–516.
102. Cox IM, Campbell MJ, Dowson D. Red blood cell magnesium and chronic fatigue syndrome. *Lancet* 1991; **337**: 757–760.
103. Monks J. Interpretation of subjective measures in a clinical trial of hyperbaric oxygen therapy for multiple sclerosis. *J Psychosom Res* 1988; **32**: 365–372.
104. O'Brien BJ, Banner NR, Gibson S, Yacoub MH. The Nottingham Health Profile as a measure of quality of life following combined heart and lung transplantation. *J Epid Commun Health* 1988; **42**: 232–234.
105. Parr G, Darekar B, Fletcher A, Bulpitt CJ. Joint pain and quality of life: results of a randomized trial. *Br J Clin Pharmac* 1989; **27**: 235–242.
106. Allen JK, Becker DM, Swank RT. Factors related to functional status after coronary artery bypass surgery. *Heart and Lung* 1990; **19**: 337–343.
107. Allen JK, Fitzgerald ST, Swank RT, Becker DM. Functional status after coronary artery bypass grafting and percutaneous transluminal coronary angioplasty. *Am J Cardiol* 1990; **66**: 921–925.
108. Lancaster TR, Singer DE, Sheehan MA, *et al.* The impact of long-term warfarin therapy on quality of life: evidence from a randomized trial. *Arch Intern Med* 1991; **151**: 1944–1949.
109. Wiklund I, Romanus C. A comparison of quality of life before and after arthroplasty in patients who had arthritis of the hip joint. *J Bone Joint Surg* 1991; **73-A**: 765–769.
110. Lansky D, Butler JBJ, Waller FT. Using health status measures in the hospital setting: from acute care to outcomes management. *Med Care* 1992; **30(Suppl)**: MS57–MS73.
111. Messick S. The once and future issues of validity: assessing the meaning and consequences of measurement. In: Wainer H, Braun H, eds. *Test Validity.* Hillsdale, NJ: Lawrence Erlbaum, 1988: 33–45.
112. Berwick DM, Murphy JM, Goldman PA, *et al.* Performance of a five-item mental health screening test. *Med Care* 1991; **29**, 169–176.
113. Street RL, Gold WR, McDowell T. Using health status measures in medical consultations. *Med Care* 1994; **32**: 732–744.
114. Schor EL, Lerner DJ, Malspeis S. Physician's assessment of functional health status and well-being: the patient's perspective. *Arch Intern Med* 1995; **155**: 309–314.
115. Health and Public Policy Committee, American College of Physicians. Comprehensive functional assess-

ment for elderly patients. *Ann Intern Med* 1988; **107**: 70–72.
116. Calkins DR, Rubenstein LV, Cleary PD, Jette AM, Brook RH, Delbanco TL. The Functional Status Questionnaire: a controlled trial in a hospital-based practice. *Clin Res* 1986; **34**: 359A.
117. Rubenstein LV, Calkins DR, Young RT, *et al.* Improving patient functional status: can questionnaires help? *Clin Res* 1986; **34**: 835A.
118. Rubenstein LV, Calkins DR, Young RT, *et al.* Improving patient function: a randomized trial of functional disability screening. *Ann Intern Med* 1989; **111**: 836–842.
119 Kazis LE, Callahan LF, Meenan RF, Pincus T. Health status reports in the care of patients with rhematoid arthritis. *J Clin Epid* 190; **43**: 1243–1253.
120. Krumholz HM, McHorney CA, Clark L, *et al.* Changes in health-related quality of life after elective percutaneous coronary revascularization. Manuscript under review.
121. Parkerson GR. February 16, 1995. [Personal communication]
122. Cronbach LJ. *Essentials of Psychological Testing.* New York: Harper and Brothers, 1949.
123. Flanagan JC. Units, scores, and norms. In: Fuchs VR, ed. *Essays in the Economics of Health and Medical Care.* New York: National Bureau of Economic Research, 1972: 695–763.
124. Applegate WB. Use of assessment instruments in clinical settings. *JAGS* 1987; **35**: 45–50.
125. Hays RD, Sherbourne CD, Mazel RM. The RAND 36-Item Health Survey 1.0. *Health Economics* 1993; **2**: 217–227.
126. Bozzete SA, Hays RD, Berry SH, Kanouse DE. A perceived health index for use in persons with advanced HIV disease: derivation, reliability, and validity. *Med Care* 1994; **32**: 716–731.
127. Ware JE, Kosinski M, Bayliss MS, *et al.* Comparisons of methods for the scoring and statistical analysis of the SF-36 health profile and summary measures: results from the Medical Outcomes Study. *Med Care,* 1995; in press.
128. Hambleton RK. Principles and selected applications of item response theory. In: Linn LR, ed. *Educational Measurement* (3rd edn). New York: Macmillan, 1989: 147–200.
129. McHorney CA, Haley SM, Ware JE. Evaluation of the MOS SF-36 physical functioning scale (PF-10). II. Comparison of relative precision using Likert and Rasch scoring methods. Under review.
130. Haley SM, McHorney CA, Ware JE. Evaluation of the MOS SF-36 physical functioning scale (PF-10). I. Unidimensionality and reproducibility of the Rasch Item Scale. *J Clin Epidem* 1994; **47**: 671–684.
131. Rector TS, Kubo SH, Cohn JN. Patients' self-assessment of their congestive heart failure. Part 2: Content, reliability and validity of a new measure. The Minnesota Living with Heart Failure Questionnaire. *Heart Failure* 1987: 198–209.
132. Tugwell P, Bombardier C, Buchanan WW, *et al.* The MACTAR Patient Preference Disability Questionnaire—An individualized functional priority approach for assessing improvement in physical disability in clinical trials in rheumatoid arthritis. *J Rheum* 1987; **14**: 446–451.
133. Mitchell A, Guyatt G, Singer J. Quality of life in patients with inflammatory bowel disease. *J Clin Gastroenterol* 1988; **10**: 306–310.
134. Bunderson CV, Inouye DK, Olsen JB. The four generations of computerized educational measurement. In: Linn LR, ed. *Educational Measurement* (3rd edition). New York: Macmillan, 1989: 367–407.
135. Koran LM. The reliability of clinical methods, data and judgements (first of two parts). *NEJM* 1975; **293(13)**: 642–646.
136. Koran LM. The reliability of clinical methods, data and judgements (second of two parts). *NEJM* 1975; **293(14)**: 695–701.
137. Helmstadter GC. *Principles of Psychological Measurement.* New York: Appleton-Century-Crofts, 1964.
138. Sox HC, Blatt MA, Higgins MC, Marton KI. *Medical Decision Making.* Boston: Butterworths, 1988.
139. Panzer RJ, Black ER, Griner PF. *Diagnostic Strategies for Common Medical Problems.* Philadelphia: American College of Physicians, 1991.
140. Hubbell AF, Waitzkin H, Rodriguez FI. Functional status and financial barriers to care among the poor. *Southern Med J* 1990; **83**: 548–550.
141. Hunt SM, McKenna SP, Williams J. Reliability of a population tool for measuring perceived health problems: a study of patients with osteoarthritis. *J Epid Commun Health* 1981; **35**: 297–300.

Appendix

Table A.1. Floor and ceiling effects for the NHP, n = 353

	% Floor	% Ceiling
Physical mobility	0	74
Pain	1	78
Energy	11	62
Social isolation	1	78
Emotional reactions	1	48
Sleep	1	62

Source: Anderson, Sullivan, and Usherwood, 1990[67]

Table A.2. Floor and ceiling effects for the COOP charts, n = 1,753

	% Floor	% Ceiling
Physical	5	32
Work	1	55
Pain	6	27
Overall	3	13
Social	2	66
Emotional	3	38
Social support	4	63
Quality of life	1	24
Change in health	1	12

Source: Medical Outcomes Study: all patients who completed baseline COOP Charts in person

Table A.3. Floor and ceiling effects for the DUKE, $n = 683$

	% Floor	% Ceiling
Physical	1	8
Mental health	1	12
Social	0	4
General	0	1
Perceived health	4	46
Anxiety	0	7
Depression	1	9
Pain	19	27
Disability	5	72
Self-esteem	0	16

Source: Parkerson GR, 1994.[68]

Table A.4. Floor and ceiling effects for the SF-36 survey, $n = 3,445$

	% Floor	% Ceiling
Physical functioning	1	19
Role—physical	24	37
Pain	1	18
General health perceptions	0	1
Vitality	1	1
Social functioning	1	46
Role—emotional	18	56
Mental health	0	4

Source: McHorney CM, Ware JE, Lu JFR, *et al.*, 1994.[69]

Table A.5. Reliability estimates and standard errors of measurement for the FSQ

	Internal-consistency reliability*	95% CI of the SEM**	Test–retest reliability†	95% CI of the SEM
ADLs	0.79	18.1	†	†
IADLs	0.82	22.9	†	†
Work performance	0.65	17.2	†	†
Social activity	0.65	41.0	†	†
Mental health††	0.81	16.7	†	†
Quality of interaction	0.64	19.5	†	†
Employment status	§	§	†	†
Frequency of social contact	§	§	†	†
Bed days	§	§	†	†
Restricted days	§	§	†	†
Sexual relationships	§	§	†	†
Feeling about health	§	§	†	†

* Source: Jette *et al.*, 1986[48]
** Standard deviations not reported by Jette *et al.*, 1986.[48] For illustrative purposes only, standard deviations used to calculate SEM were derived from Hubbell, Waitzkin, and Rodriguez (1990).[140]
† Data not published on test-retest reliability
†† Identical to the SF-36 mental health scale
§ Not applicable; single-item measures

Table A.6. Reliability estimates and standard errors of measurement for the NHP

	Internal-consistency reliability*	95% CI of the SEM**	Test–retest reliability†	95% CI of the SEM
Physical mobility	0.39	32.1	0.85	15.9
Pain	0.72	28.9	0.79	25.1
Energy	0.57	41.5	0.77	30.4
Social isolation	0.34	37.7	0.78	21.8
Sleep	0.68	31.8	0.85	21.8
Emotional reactions	0.81	23.0	0.80	23.6

* Source: Wiklund, Romanus, and Hunt, 1988[81]
** Standard deviations not reported by Wiklund, Romanus, and Hunt, 1988.[81] For illustrative purposes only, standard deviations used to calculate SEM were derived from Jenkinson, Fitzpatrick, and Argyle, 1988[1]
† Source: Hunt, McKenna, and Williams, 1981[141]

Table A.7. Reliability estimates and standard errors of measurement for the COOP charts

	Alternate-forms reliability*	95% CI of the SEM	Test–retest reliability**	95% CI of the SEM
Physical	0.37	45.0	0.83	32.3
Role	0.43	35.7	0.88	20.4
Pain	0.41	46.2	0.74	37.5
Overall	0.41	36.7	0.71	31.7
Social	0.24	39.9	0.64	38.2
Emotional	0.54	36.6	0.66	34.3
Social support	†	†	0.46	50.4
Quality of life	†	†	0.42	37.3
Change in health	†	†	†	†

* Source: McHorney *et al.*, 1992[92]
** Source: Nelson *et al.*, 1990[85]
† Reliability estimates not available

Table A.8. Reliability estimates and standard errors of measurement for the DUKE

	Internal-consistency reliability*	95% CI of the SEM	Test–retest reliability*	95% CI of the SEM
Physical health	0.67	25.0	0.75	21.8
Social health	0.55	23.1	0.57	22.6
Mental health	0.68	22.8	0.70	22.1
Self-esteem	0.64	21.5	0.78	16.8
Anxiety	0.60	22.7	0.62	22.1
Depression	0.65	24.8	0.68	23.7
General health	0.78	14.2	0.78	14.2
Pain	**	**	0.41	50.3
Disability	**	**	0.30	46.4
Perceived health	**	**	0.56	34.3

* Source: Parkerson, Broadhead, and Tse, 1990[23]
** Single-item measure; reliability and SEM not computed

Table A.9. Reliability estimates and standard errors of measurement for the SF-36 scales

	Internal-consistency reliability*	95% CI of the SEM	Test–retest reliability*	95% CI of the SEM
Physical functioning	0.93	13.8	0.81	22.7
Role—physical	0.84	31.8	0.69	44.2
Pain	0.82	20.9	0.78	23.1
General health perceptions	0.78	19.5	0.80	18.6
Vitality				
Social functioning	0.87	15.6	0.80	19.3
Role—emotional	0.85	18.5	0.60	30.2
Mental health	0.83	32.1	0.63	47.3
	0.90	13.1	0.75	20.7

* Source: McHorney *et al.*, 1994[69]
** Source: Brazier *et al.*, 1992[70]