Published in Health Survey Research Methods, Conference Proceedings, by Floyd J. Fowler, Jr., Ph.D. (Ed.), National Center for Health Services Research and Health Care Technology Assessment, September, 1989, pgs. 13-21.

Comparison of Responses to Similar Questions in Health Surveys

John P. Anderson, Robert M. Kaplan, and Margaret DeBon

Introduction

Over the last several decades, recognition of the need for sensitive indicators of health status and quality of life has increased. This need is apparent for several reasons. First, current health indicators are inadequate for capturing many of the health status variables that are associated with the need for health care. Measures of mortality provide hard end points but ignore all of those who are alive. Fries and associates (1989) emphasize that the likelihood of extending current life expectancy for adults is very small. Thus, there is remarkably little evidence that major medical and preventive interventions that apply to those who have survived their first years of life actually make people live longer. Yet, as Fries and colleagues (1989) have argued, substantial public health benefits may be achieved by compressing morbidity toward the end of the life cycle. Evaluating these interventions will require more sensitive measures of health outcome. Current data from the National Health Interview Survey (NHIS) provide information that only a minority of the U.S. population are, by their standards, in ill health. In 1985, for example, 90 percent of the U.S. population was reported to be in excellent, very good, or good health. A substantial majority (86 percent) reported no activity limitations (Dawson and Adams, 1987).

This paper suggests that many current techniques for evaluating health status and quality of life are insensitive for detecting important variations in health status. Specifically, it is argued that variations in the experience of what we have come to call Symptom/Problem Complexes (CPX) are, by patient-citizen preference standards, highly important to how they come to evaluate their health status. This implies that approaches that rely exclusively on dysfunction are seriously deficient in their sensitivity to important dimensions in measuring health status. Data from several studies are presented to suggest that seemingly minor variations in the wording of survey questions can produce significant differences in the estimates of the extent of dysfunction in and overall health status of populations.

Although a growing number of studies now incorporate health-related quality of life measures, there has been a strong emphasis on cost savings and time efficiency. Self-administered questionnaires are frequently assumed to be the better alternative because they are cheap and easy. Over the last two decades, our group has worked toward the development of a General Health Policy Model (Kaplan and Anderson, 1988). One of the objectives of this line of research is the development of a valid and reliable questionnaire for assessing health-related quality of life. Several studies have identified problems, particularly in the underreporting of dysfunction (Reynolds & associates, 1974; Stewart & associates, 1981). In several of our studies, both self-administered and interviewer-administered questionnaires were given to the same respondents. The results are of interest not only because of mode of administration but because they provide information on type of question. This paper summarizes three studies from our current research program. All of these studies use the Quality of Well-being (QWB) scale and instrument, which will now be briefly described.

Quality of Well-being Scale

The QWB scale combines preference-weighted measures of symptoms and functioning to provide a numerical point-in-time expression of well-being, which ranges from zero (0) for death, to one (1.0) for asymptomatic

John P. Anderson, Robert M. Kaplan, and Margaret DeBon are with Division of Health Care Sciences, Department of Community and Family Medicine, University of California, San Diego.

This research was supported in part by Grant No. R18 HS 05617 from the National Center for Health Services Research and Health Care Technology Assessment and Grant AR 33489 from the National Institutes of Health.

Table 1. List of Quality of Well-being Scale Symptom/Problem Complexes (CPX) with calculating weights

CPX no.	CPX description	Weights	CPX no.	CPX description	Weights
1	Death (not on respondent's card)	−0.727	13	Headache, or dizziness, or ringing in ears, or spells of feeling hot, or nervous, or shaky	−.244
2	Loss of consciousness such as seizure (fits), fainting, or coma (out cold or knocked out)	−.407	14	Burning or itching rash on large areas of face, body, arms, or legs	−.240
3	Burn over large areas of face, body, arms, or legs	−.367	15	Trouble talking, such as lisp, stuttering, hoarseness, or inability to speak	−.237
4	Pain, bleeding, itching, or discharge (drainage) from sexual organs—does not include normal menstrual (monthly) bleeding	−.349	16	Pain or discomfort in one or both eyes (such as burning or itching) or any trouble seeing after correction	−.230
5	Trouble learning, remembering, or thinking clearly	−.340	×17	Overweight *or underweight* for age and height or skin defect of face, body, arms or legs, such as scars, pimples, warts, bruises, or changes in color	−.186
6	Any combination of one or more hands, feet, arms, or legs either missing, deformed (crooked), paralyzed (unable to move) or broken—includes wearing artificial limbs or braces	−.333	18	Pain in ear, tooth, jaw, throat, lips, tongue; missing or crooked permanent teeth—includes wearing bridges or false teeth; stuffy, runny nose; any trouble hearing—includes wearing a hearing aid	−.170
7	Pain, stiffness, weakness, numbness, or other discomfort in chest, stomach (including hernia or rupture), side, neck, back, hips, or any joints of hands, feet, arms or legs	−.299	19	Taking medication or staying on a prescribed diet for health reasons	−.144
8	Pain, burning, bleeding, itching, or other difficulty with rectum, bowel movements, or urination (passing water)	−.292	20	Wore eyeglasses or contact lenses	−.101
9	Sick or upset stomach, vomiting or loose bowel movements, with or without fever, chills, or aching all over	−.290	21	Breathing smog or unpleasant air	−.101
10	General tiredness, weakness, or weight loss	−.259	22	No symptoms or problem (not on respondent's card)	−.000
11	Cough, wheezing, or shortness of breath with or without fever, chills, or aching all over	−.257	23	Standard symptom/problem (not on respondent's card)	−.257
12	Spells of feeling upset, being depressed, or of crying	−.257	×24	*Trouble sleeping*	−.257
			×25	*Intoxication*	−.257
			×26	*Problems with sexual interest or performance*	−.257
			×27	*Excessive worry or anxiety*	−.257

optimum functioning. Table 1 presents 25 Symptom/Problem Complexes along with their preference weights. Use of this CPX list does not require any assumptions about the intensity or duration of symptoms and problems nor the underlying pathology, if any. This measure simply indicates that symptoms are present or absent on a given day.

Quality of Well-being also involves three scales of function: Mobility (MOB), Physical Activity (PAC), and Social Activity (SAC). Each step on these scales has its own associated preference weight. These are reported in Table 2, along with the single-day QWB calculating formula (formula 1). In the General Health Policy Model, QWB inputs are integrated with terms for the number of people affected and the duration of time affected to produce the output expression of Well-years (formula 2).

Study I: Evaluation of Self-Administered QWB Items

Method

Data from this analysis come from a household interview survey of a sample of 1,324 subjects. These subjects included 866 randomly selected respondents, 369 randomly selected children, and 89 persons with a physical dysfunction who were selected on the basis of responses to screening questions. Seventy-seven percent of those initially contacted completed the study.

Figure 1 characterizes the types of data available. Each respondent answered all questions relevant to functioning on the three scales in a self-report mode. In addition, respondents were assessed by a trained interviewer. The order of presentation was counterbalanced to control for order effects. In each case, questions were

Table 2. Quality of Well-being General Health Policy Model elements and calculating formulas (function scales, with step definitions and calculating weights)

Step No.	Step definition	Weight
	Mobility scale (MOB)	
5	No limitation for health reasons	−0.000
4	Did not drive a car, health related; did not ride in a car as usual for age (younger than 15 yr), health related, and/or did not use public transportation, health related; or had or would have used more help than usual for age to use public transportation, health related	−.062
2	In hospital, health related	−.090
	Physical activity scale (PAC)	
4	No limitations for health reasons	−.000
3	In wheelchair, moved or controlled movement of wheelchair without help from someone else; or had trouble or did not try to lift, stoop, bend over, or use stairs or inclines, health related; and/or limped, used a cane, crutches, or walker, health related; and/or had any other physical limitation in walking, or did not try to walk as far or as fast as others the same age are able, health related	−.060
1	In wheelchair, did not move or control the movement of wheelchair without help from someone else, or in bed, chair, or couch for most or all of the day, health related	−.077
	Social activity scale (SAC)	
5	No limitations for health reasons	−.000
4	Limited in other (for example, recreational) role activity, health related	−.061
3	Limited in major (primary) role activity, health related, but did perform self-care activities	−.061
2	Performed no major role activity, health related, but did perform self-care activities	−.061
1	Performed no major role actvity, health related, and did not perform or had more help than usual in performance of one or more self-care activities, health related	−.106

Calculating formulas

Formula 1. Point-in-time well-being score for an individual (W):

$$W = 1 + (\text{CPX}wt) + (\text{MOB}wt) + (\text{PAC}wt) + (\text{SAC}wt)$$

where wt is the preference-weighted measure for each factor and CPX is Symptom/Problem complex. For example, the W score for a person with the following description profile may be calculated for one day as:

CPX-11	Cough, wheezing, or shortness of breath, with or without fever, chills, or aching all over	−0.257
MOB-5	No limitations	−.000
PAC-1	In bed, chair, or couch for more or all of the day, health related	−.077
SAC-2	Performed no major role activity, health related, but did perform self-care	−.061

$$W = 1 + (-.257) + (-.000) + (-.077) + (-.061) = .605$$

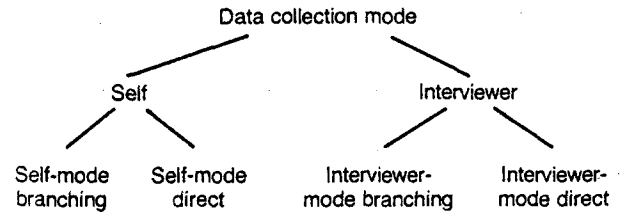
Formula 2. Well-years (WY) as an output measure:

$$WY = \text{No. of Person} \times (\text{CPX}wt + \text{MOB}wt + \text{PAC}wt + \text{SAC}wt) \times \text{Time}$$

presented in both a branching and direct mode. In the branching mode, the respondents answered an algorithmic series of closed questions and branching follow-up probes. First, questions asked whether the subjects actually performed a specific activity. If they did not, a probe question was used to determine the reasons for nonperformance. Both yes and no answers were probed in fuller detail. Strict criteria were used to code whether or not reasons for nonperformance were related to health. The questions were designed for either interviewer or self administration. Examples of the branching and direct questions for the self-administered questions for the mobility portion of the Quality of Well-being are given in Table 3.

In the self mode, the respondents were directed to read definitions for all steps in the scales. The respondent then reported to the interviewer the number of the step on each scale that best described themselves and/or the other subjects for whom they reported. The self-read definitions required the respondent to interpret whether any nonperformance of activities was due to

Figure 1. Categories of evidence



SOURCE: Anderson & associates (1986)

3. Self-mode direct and branching question patterns by study from mobility scale

Initial Survey, Direct Mode Card B (Mobility Scale)	Follow-up Survey, Branching Mode Card II (Mobility Scale, Over 16)
ı special unit of a hospital such as an operating or ›very room, intensive care unit, incubator, isolation d, for any part of a day ı hospital, nursing home, mental hospital, home for ırded as a patient ded help to go outside, or stayed inside all day for lth reasons ıld go outside without help, but could not drive and/or ld not use public transportation without help from ther person. (For a child: needed more help to travel ı usual for age.) ɜ to both drive and use public transportation (bus, ı, etc.) without help. (For a child: able to travel as ıl for age.)	In each category choose the numbers* that A. Spent any part of the day or night as a bed patient in a hospital, nursing home, mental institution, home for the retarded, or similar place. A1. Yes A2. No B. Driving B1. Drove car (or motor vehicle) B2. Did not drive, for health reasons B3. Did not drive, for reasons not related to health C. Public Transportation Use bus, train, plane or subway C1 Without help from anyone else C2 With help from another person for health reasons Did not use bus, train, plane, or, subway. C3 For health reasons C4 For reasons not related to health.

s in self-mode branching do not correspond to scale steps.
:: Anderson and associates, 1986

related reasons. Although the respondents were ›d to read the items, they were not requested to the information on their own. Thus, test of the ıde should have created the most favorable con- for self-administration.

ng long interviews a variety of other questions servations were made. For example, the inter- also engaged the respondent in open-ended dis-, completed interviewer notes, and tape recorded terview. This information was used to estimate / well the questions were being understood; (2) ısely the respondents understanding matched the ɜ of the question; and (3) how closely the cate- ıswers in both types of administration matched ıal situation. When a discrepancy between the ɜ of categoric responses was observed, all of this ıtion was systematically studied to estimate the tely true classification for the respondent.

l analysis suggested that correlations between 'erent modes of administration were very high. they tended to be .98 or higher! Even for those ents who were highly dysfunctional, correlation ı modes of administration tended to be .90 or Despite high correlations between overall QWB ›r the different modes of administration, there ubstantial number of inconsistencies in function ıtion between these modes. To evaluate these, analysis was conducted (Anderson and associ- 6; 1988). Table 4 summarizes the method used sensitivity and specificity of the different forms. h these methods are common, the table includes a few uncommon terms. Each of the respondents are classified into one of five categories. These include (a) report of dysfunction when there is, indeed, dysfunction; (b) reports of dysfunction when there is no dysfunction; (c) reports of no dysfunction when, indeed, there is dysfunction; and (d) report of no dysfunction when there is no true dysfunction. The final category (e) is for people who correctly report they are dysfunctional but are placed in the wrong dysfunctional category. In addition to calculating sensitivity and specificity by standard methods, we offer new concepts for predictive value of dysfunctional and predictive value of function. The predictive value of dysfunction is the ratio of those who report the correct dysfunctional category over all those reporting dysfunction. The predictive value of functional reporting is the ratio accurately reporting function over all reports of functioning. Table 5 displays the validity characteristics for both modes of administration. The two modes of administration differ dramatically in accurately classifying dysfunction when the dysfunction state is compared to actual dysfunction. These errors are reflected in the sensitivity of the measure. Analysis suggested that the sensitivity of the PAC scale was .45 in the self-administered version. In other words, only 45 percent of the actual dysfunction was captured. In contrast, the interviewer-administered version accurately classified 86 percent. The predictive value of dysfunction was also low in the self-administered versions and considerably higher in the interviewing administered version. Specificity was high for both modes of administration.

In the early days of development of the Quality of Well-being scale, a self-administered questionnaire was seen as highly desirable. The interviewer mode was cho-

leasurement categories and validity characteristics modified for multiple state analysis

Measurement categories for multiple states

	ACTUAL DYSFUNCTION	ACTUAL (FULL) FUNCTION	
ysfunction	(a) Correctly classified dysfunction (e) Misclassified dysfunction	(b) False dysfunction	Total reported dysfunction (= a + b + e)
full) function	(c) False function	(d) Full function	Total reported (full) function (= c + d)
	Total dysfunction (= a + c + e)	Total actual (full) function (= b + d)	

Validity characteristics modified for multiple states

$$\text{Sensitivity} = \frac{\text{Correctly classified dysfunction}}{\text{Total actual dysfunction}} = \frac{a}{a + c + e}$$

$$\text{Predictive value dysfunctional} = \frac{\text{Correctly classified dysfunction}}{\text{Total reported dysfunction}} = \frac{a}{a + b + e}$$

$$\text{Specificity} = \frac{\text{Full function}}{\text{Total actual (full) function}} = \frac{d}{b + d}$$

$$\text{Predictive value functional} = \frac{\text{Full function}}{\text{Total reported (full) function}} = \frac{d}{c + d}$$

derson & associates, 1986

old standard against which to evaluate the less self-administered mode. However, as these est, there may be serious problems with the of the self-administered mode. Other studies reported problems in detecting limitations dministered scales. For example, the inability imitations with single closed-ended questions, acknowledgment of limitations in the same was also reported in data from the National terview Survey (Cannell and others 1977). In Health Insurance Study, there was missing or nt information on 37 percent of the functional s reported. We observed problems with the ting of functional limitations in 28.7 percent of with self-administered questionnaires.

re many potential explanations for these find- is that respondents misunderstand questions f-administered forms. This problem becomes the complexity of the questions increases. branching questions, sensitivity will decrease. ability of a trained interviewer allows the de- on of actual performance versus nonperform- tivities. In addition, nonperformance of activ- be evaluated as related to health or for reasons. Further, an interviewer can assess ic days for which there was a problem. Evi- s developed that sequential branching ques- require an interviewer can reliably penetrate exity of quality of life and dysfunctional states. o noted that a very high proportion of the n experiences at least some minor dysfunction m on a particular day. In the community sur- xample, interviewer-administered question- h branching patterns identify only about 11 f the population as completely functional and asymptomatic, by comparison to the NHIS identification of 86 percent of the population as completely functional, without regard to experience of symptoms. Clearly, this wide difference means variations in what has been called

Table 5. Validity characteristics by data collection method and format

Scale Validity characteristics	Initial survey · direct	Follow-up survey combined days branching
Self mode, direct and branching		
Mobility		
Sensitivity	0.68	0.66
Predictive value dysfunctional	.56	.41
Specificity	.98	.95
Predictive value functional	.99	.98
Physical and social activity		
Sensitivity	.45	.61
Predictive value dysfunctional	.59	.73
Specificity	.99	.99
Predictive value functional	.94	.96
Interviewer mode, branching		
All scales combined		
Sensitivity	.89	.86
Predictive value dysfunctional	.93	.91
Specificity	.99	.99
Predictive value functional	.99	.99

SOURCE: Adapted from Anderson and associates, 1986

high-level wellness are absolutely critical for sensitive and accurate measurement of health status.

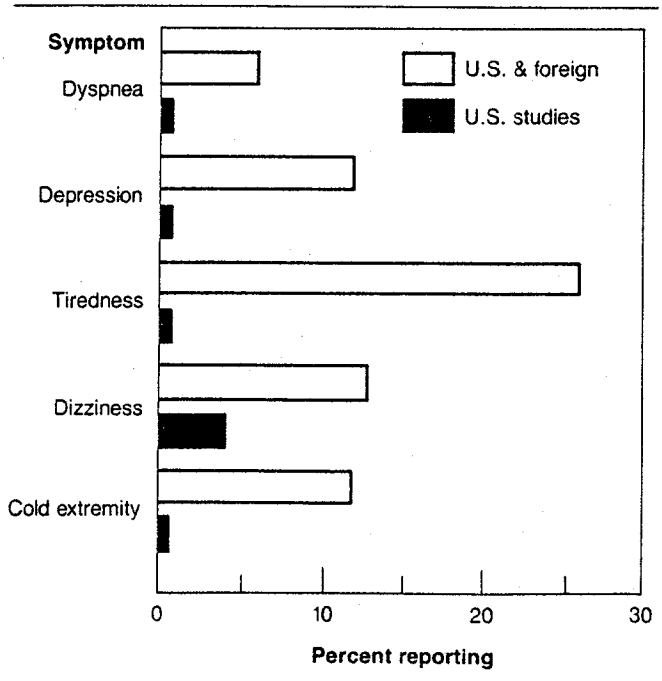
This work on structured interviews about function led to questions about the value of self-report symptom inventories. That issue was investigated in Study II.

Study II: An Experiment on Symptom Reporting

Clinical studies with the same experimental design show considerable variability in the effects of treatment on symptoms. This may be of particular concern in studies of drug side effects. Consider, for example, Figure 2. The data for this figure were taken from an advertisement for Atenolol that was published in the *Journal of the American Medical Association*. In the very small print of the advertisement, side effects of the medication were reported. The ad separated data from U.S. studies and U.S. plus foreign studies. As the figure suggests, side effects in the American studies are quite rare. Yet the very same side effects are actually quite common in U.S. plus foreign studies. Consider, for example, tiredness which occurs in 0.6 percent of U.S. studies, but 27 percent of U.S. plus foreign studies. It is presumed that U.S. plus foreign studies are combined in order to dilute what may be very common side effects in the foreign studies. Similar results are apparent for dyspnea, depression, and other symptoms.

Why do these results from studies of the same product produce such different results? One explanation is in the way that symptoms were assessed. Typically, U.S. drug studies ask about only a small number of symptoms. Then patients are asked in a free format if they have any other symptoms. In European studies, there is a systematic symptom-by-symptom inquiry. In our work on the Quality of Well-being scale, respondents are presented with a list of symptoms that is meant to be exhaustive. Then they are asked to identify which symptom complexes they have experienced for each day over the last 6 days. An alternative procedure would be to have the interviewer read each individual symptom and ask whether that symptom had been experienced (Eakin, Kaplan, & Ganiats, 1989). These formats may lead to differential report rates.

Figure 2. Symptoms associated with use of atenolol in U.S. and U.S. and foreign studies



Subjects

The participants in the study were 82 adults who were being cared for by the family medicine practice at the University of California, San Diego. All were followed by their physicians for routine health problems or other conditions that do not require the attention of a specialist.

Procedure

The patients were randomly assigned to one of two groups. Group 1 was given the standard instruction which is:

> For most of these questions, I'll be asking about the past six days, that is, from (day/date) through (day/date). First, I would like to ask you about any health problems you might have had. Please look at this list one at a time and tell me the number of all the items that you had at any time during the past six days. Don't worry about how important or serious the problem was; if it was present at all in the last six days, please give me the number. Were there any health problems not on the list that you had at any time during the past six days?

For Group 2, the interviewer proceeded through the symptom problem list and requested the patients to report whether they experienced each item. The data analysis involved a *t*-test comparing the mean number of symptoms reported for each of the two conditions.

Results

The group receiving the standard instruction reported an average of 2.64 symptoms per day while the group receiving the item by item instruction reported an average of 2.86 symptoms per day. These differences were not statistically significant ($p = 0.55$).

Study III: Comparison of Similar Items on Different Standardized Questionnaires

The third study considers a somewhat different question. In this, we compared responses to very similar items that were developed for different standardized questionnaires. Specifically, we compared responses to questions on the Quality of Well-being scale with items on an arthritis-specific measure known as the Arthritis Impact Measurement Scale (AIMS). There were several reasons for these comparisons. First, the Arthritis Impact Measurement Scale is commonly used in arthritis

Table 6. Comparison of similar items in Quality of Well-being and Arthritis Impact Measurement Scale

Items	Percent agreement	Percent QWB dysfunction	Percent AIMS dysfunction
AIMS 1. When you travel around your community, does someone have to assist you because of your health? MOB 2. On (day/date) were there reasons related in any way to your health that you did not (drive a car/ride in a car)? What were the reasons? On (day/date) (did you/would you) use more help from someone else than usual for your age?	98	0	2
AIMS 2. Are you able to use public transportation? MOB 3. On which of the past 6 days, if any, did you use public transportation, such as a bus, plane, train, or trolley? On (day/date) were there reasons related in any way to your health that you did not use public transportation? On (day/date) did you use, or would you have used, more help from someone else than usual for your age to take public transportation?	92	0/10*	8
AIMS 4. Are you in bed or a chair for most or all of the day because of your health? PAC 3. On which of the past 6 days, if any, did you spend most or all of the day in any type of chair or couch?	85	11	4
AIMS 6. Do you have any trouble either walking several blocks or climbing a few flights of stairs because of your health? PAC 6. On which of the past 6 days, if any, did you have any other physical limitation or not try to walk as far or as fast as most persons your age are able?	69	10	21
AIMS 7. Do you have trouble bending, lifting, or stooping because of your health? PAC 4. On which of the past 6 days, if any, did you have trouble, or not try, to lift, stoop, bend over, or use stairs or inclines?	76	9	15
AIMS 8. Do you have any trouble either walking one block or climbing one flight of stairs because of your health? PAC 6. On which of the past 6 days, if any, did you have other physical limitation or not try to walk as far or as fast as most persons your age are able?	67	25	8
AIMS 9. Are you unable to walk unless you are assisted by another person or by a cane, crutches, artificial limbs, or braces? PAC 5. On which of the past 6 days, if any, did you limp or use a cane, crutches, or walker?	64	36	0
AIMS 15. If you had the necessary transportation, could you go shopping for groceries or clothes? SAC 1B. If you had worked (or did work) on (day/date), were you limited in the amount or kind of work done, such as using special working aids, not doing certain tasks, taking special rest periods, or working only part of the day?	87	9	4
AIMS 26. When you bathe, either a sponge bath, tub, or shower, how much help do you need? SFC 4. Did not take bath for health reasons or had help to take bath (getting in or out of tub or shower, washing all parts of the body, etc.)	95	2	3
AIMS 27. How much help do you need in getting dressed? SFC 1. Did not dress for health reasons, or had help to dress (tying shoes, buttoning shirt, blouse, coat, etc.).	89	7	4
AIMS 28. How much help do you need to use the toilet? SFC 3. Did not use toilet for health reasons (e.g., bedpan) or had help to use toilet (getting on or off the seat, cleaning with tissues, etc.)	99	1	0
AIMS 31. During the past month how often have you had severe pain from your arthritis? CPX 7. Pain, stiffness, weakness, numbness, or other discomfort in chest, stomach, side, neck, back, hips, or any joints of hand, feet, arms, or legs.	91	9	0
AIMS 38. During the past month, how much of the time have you been in low or very low spirits? CPX 12. Spells of feeling upset, depressed, or crying.	47	1	52

research. It is believed to be more sensitive to clinical changes in arthritis patients because the items are arthritis specific. However, the Arthritis Impact Measurement Scale is often self-administered and therefore includes many of the same potential difficulties as do other self-administered questionnaires.

A second reason for conducting this analysis is that there is growing interest in imputing scores for one mea-

sure retrospectively from data collected using a different questionnaire (Erickson & associates, 1988, 1989). For example, the National Health Interview Survey does not include sensitive measures that can be used for quality-of-life evaluations. In addition, many policy analyses require data that are not available in the standard National Center for Health Statistics (NCHS) questionnaires. Nevertheless, items on the national survey are quite similar to those used in some quality-of-life measures. Thus, there is interest in imputing the more sensitive quality-of-life measures from responses given in the national surveys. These imputations make the assumption that responses from one measure can be accurately predicted from responses on another measure. Study III tests this assumption.

Method

The subjects were 92 adults with musculoskeletal diseases treated by the Scripps Clinic and Research Foundation. The rationale for selecting only patients with musculoskeletal disorders was that the Arthritis Impact Measurement Scale instrument was only appropriate to them. Using a nonhealthy population maximizes the number of estimated dysfunctional states in the population. The Quality of Well-being and Arthritis Impact Measurement Scale questionnaires were both administered by a trained interviewer during regular clinic visits. Table 6 shows the items in the two scales that are used for comparison. In addition, the table shows the percentage of patients for which there was agreement, defined as reporting a problem on both or neither of the items. Table 6 also shows the percentage of cases where only the Quality of Well-being questionnaire or only the Arthritis Impact Measurement Scale questionnaire detected health problems. As Table 6 suggests, there tended to be high agreement between the two measures for most items. Among 13 items with similar wording, the average agreement score was 82 percent. The Quality of Well-being detected more problems in eight items whereas the Arthritis Impact Measurement Scale detected more problems in 5 cases.

The cases of large discrepancy between the Arthritis Impact Measurement Scale and Quality of Well-being typically compared questions in which there were subtle differences in wording. For example, there was a large difference between Arthritis Impact Measurement Scale 9 and Physical Activity 5 from the Quality of Well-being. One difference in these questions is that the Quality of Well-being items ask about limping, and 25 patients reported a limp. The Arthritis Impact Measurement Scale does not inquire about limping.

Another disturbing discrepancy is between the AIMS question on depression and the Quality of Well-being symptom-problem for depression. A remarkable 78 percent of the arthritis patients reported depression on at least one measure. Among the 47 percent for which there was agreement, 55 percent reported depression on both scales whereas 45 percent reported it on neither. However, the Arthritis Impact Measurement Scale was much more likely to pick up depression than was the Quality of Well-being. Although it is assumed that both items will capture depression, the Arthritis Impact Measurement Scale item assumes that people experience depression and asks for how much time in the last month they were depressed. The Quality of Well-being item asks about the last 6 days only and imbeds depression within a list of physical symptoms and problems.

Discussion

This paper reviews three different studies on alternative methods for posing the same issues to survey respondents. In all three studies, trained interviewers administered different forms of similar questions, so the interviewer factor was held constant. However, in each study one form of the question was designed for self-administration. On the basis of these studies, some general conclusions might be offered. These include:

1. Interviewer-administered questions typically detect higher rates of dysfunction. There is reason to believe that these higher rates are indeed true rates of dysfunction.
2. Although correlations between self-administered and interviewer-administered questionnaires may be high, these high correlations are dominated by variability in dysfunction within the population. The issue of sensitivity is often overlooked. Highly sensitive instruments are required to capture minor variation within specific subpopulations.
3. Embedding mental health symptoms, such as depression, within the context of physical health questions may lead to underreporting. This issue needs further study.
4. The consequences of failing to have adequate sensitivity are that health status is overestimated for a population. A related problem relevant to clinical trials is that side effects of treatments are often overlooked. In fact, there may be incentives in some trials to ignore adverse drug effects. This can be accomplished most easily by using insensitive measures of health outcome.

Establishment of a laboratory for methodological studies in health-status assessment is just beginning. These studies are very preliminary. They have small sample sizes with insufficient statistical power to answer many questions. However, this is a promising line of research that will ultimately produce more valid and reliable measures of health status and health-related quality of life. These measures may have significant ben efits for health services research, policy analysis, and assessment of outcomes in clinical trials.

References

Anderson, J. P., Bush, J. W., & Berry, C. C. (1986). Classifying function for health outcome and quality-of-life evaluation: Self-versus interviewer modes. *Medical Care*, 24, 454.

Anderson J. P., Bush J. W., & Berry C. C. (1988). Internal consistency analysis: A method for studying the accuracy of function assessment for health outcome and quality-of-life evaluation. *Journal of Epidemiology*, 41, 127.

Cannell, C. F., Marquis, K. H., & Laurent, A. (1977). The summary of studies on interviewing methodology. *Vital and Health Statistics* (Series 2, No. 69, DHEW Publication RHA77-1343). Washington, DC: U.S. Government Printing Office.

Dawson, D. A., & Adams, P. F. (1987). Current estimates from the National Health Interview Survey, United States, 1986. *Vital and Health Statistics,* Series 10. Hyattsville, MD. National Center for Health Statistics.

Eakin, E., Kaplan, R. M., & Ganiats, T. G. (1989). Methods for inquiring about symptoms in the Quality of Well-being scale. Unpublished manuscript, University of California, San Diego.

Erickson, P., Anderson, J. P., Kendall, E. A., & associates. (1988). Using retrospective data for measuring quality of life: National Health Interview Survey Data and the Quality of Well-being Scale. *Quality of Life and Cardiovascular Care,* 4, 179-184.

Erickson, P., Kendall, E. A., Anderson, J. P., & associates. (1989). Using composite health status measures to assess the nation's health. *Medical Care,* 27, 566-576.

Fries, J. F., Green, L. W., & Levine, S. (1989). Health promotion and the compression of morbidity. *Lancet* II, 481-483.

Kaplan, R. M., & Anderson, J. P. (1988). A general health policy model: Update and applications. *Health Services Research,* 23, 204-235.

Reynolds, W. J., Rushing, W. A., & Miles, D. L. (1974). The validation of a function status index. *Journal of Health and Social Behavior,* 15, 271.

Stewart, A. L., Ware, J. E., Jr., & Brook, R. H. (1981). Advances in the measurement of functional status: Construction of aggregate indexes. *Medical Care,* 19, 473.