Methods for Measuring Levels of Well-being for a Health Status Index

by Donald L. Patrick, J. W. Bush, and Milton M. Chen

Three psychological scaling procedures—category rating, magnitude estimation, and equivalence—were used to measure the levels of well-being that student and health-leader judges in 14 experimental groups associate with 50 case descriptions of function status representing the continuum from complete well-being to death. No significant differences were detected for order of method presentation, interview situation, scaling method, student vs. leader judges, or most interactions among these factors. Category rating, the simplest and apparently the most reliable of the methods, was consistent with the results of the social choices implied in the equivalence technique. The results indicate the feasibility of measuring the social values of large numbers of cases in household interview surveys.

Previous efforts to develop a health status index, a composite expression of the health status of a population, have encountered a plethora of conceptual and methodological problems [1,2]. Such an index would serve as a social indicator for comparing the health status of populations of various geographic areas, provide criteria for evaluating efforts to improve the health status of a community or target population, and help assess the projected benefits of new programs competing for limited health resources.

This study compares several methods for measuring social preferences for conditions of function status that compose the value component of health, a difficult dimension to quantify for a health status index.

Inherent in the concept of health used here are two analytically distinct dimensions: function status, or the level of well-being at a point in time, and prognosis, or the probability of transition to other function levels, more or less favorable, at future times [3,4]. Although these two dimensions are related, they require separate specification for measurement and computational purposes. Once this distinction is made, the social construct of health can be ex-

Supported by Grant No. 7-R18-HS00702 from the National Center for Health Services Research and Development and by contracts with the New York State Health Planning Commission. An earlier version of this paper was presented at the annual meeting of the American Sociological Association, New York, Aug. 27–30, 1973.

Address communications and requests for reprints to J. W. Bush, M.D., Director, Division of Health Policy and Management, Department of Community Medicine, University of California at San Diego, La Jolla, Calif. 92037.

plicitly formulated as a model that relates the value and prognostic dimensions. The value dimension of health is the level of well-being or the social preferences for levels of function on a continuum from optimal function to death. When these values have been measured, health status can be expressed precisely as the product (expected value) of the preferences associated with the levels of function at a point in time and the probabilities of transition to other levels over the remainder of a defined standard life.

Mathematically this concept may be expressed as follows:

$$H = E(F) = \frac{\sum F_j Y_j}{Y}$$

where H = health index for a population

$E(F)$ = expected function level value over a standard life defined as 100 years

F_j = social preference weight assigned to function level j

j = index for function levels = 0, 1, 2, . . . , 30

$Y_j = \sum_{t=1}^{k} \pi_{j,t}$ = expected total duration in function level j over all time periods

$\pi_{j,t}$ = proportion of time spent in function level j between time periods $t-1$ and t, derived from the product of the original distribution and the transitional probability matrixes for each time period; mathematically $\pi_{j,t}$ is the probability of being in state j in time period t

t = index of time periods = 0, 1, . . . , k, where k is the last interval before the end of the standard life

$Y = \sum_{j=0}^{30} Y_j$ = remainder of standard life

This definition formally expresses as a mathematical function the relation between function level and transition probabilities previously noted by other investigators [5,6].

Measurement of Social Preferences

Value or preference measurement is a difficult and controversial task. The purpose of this study was to evaluate several different measurement techniques to find an optimal method for scaling health values. The validity and reliability of those methods—category rating, magnitude estimation, and equivalence—were compared between individual and group interview situations, between orders of method presentation, and between students and health leaders.

In a previous study, the authors defined optimal function as conformity to social norms of well-being, including the ability to perform the daily activities usual for a person's age and social role [7]. Using classifications of norms related to social activities, mobility, and physical activity, 29 levels of function were defined that could be looked on as more or less desirable on a value continuum ranging from 1.0 for complete well-being to 0.0 for death. A matrix bounded by the 29 function levels, 5 age classes, and 42 symptom/problem complexes describes a universe of possible function status conditions [7]. While it should be possible to classify any person into one and only one of the function levels on any given day, the symptom/problem complexes are simply illustrative aggregations of factors that cause deviations from well-being.

Each function status condition consists of an age class, steps on scales of mobility, physical activity, and social activity, and a symptom/problem complex, as shown in Table 1. In the previous study, judges rated a probability sample of case descriptions drawn from each of the 29 levels. The 50 items in Table 1 had low interjudge variability and cover the range of the preference continuum; therefore they were chosen for use in the present experiment.

The preference measurement strategy in the present study is to construct a multimethod-multigroup matrix for comparing several scaling methods across different experimental groups. This approach is an adaptation of Campbell and Fiske's multitrait-multimethod and Centra's multiscale-multigroup validation procedure [8,9]. Here, however, only a single trait—the level of well-being, desirability, or social preference that society associates with different descriptions of function status—is measured. Thus convergent validity is tested by comparing several scaling methods with one another, and reliability is tested by comparing the results obtained from different groups under different experimental conditions.

The goal is to find a measurement method that will produce an equal-interval scale for the single dimension of preference that is valid and reliable, yet simple enough to be used in household interview surveys to determine social values. Previous experience in scaling both physical and nonphysical stimuli reveals that different types of measurement procedures frequently result in different scales for the same continua [10]. Methods were therefore selected from three classes of quantitative judgment models to determine the response characteristics of the well-being continuum.

Category rating, a difference or partition measure, is simple and efficient and can produce equal-interval scales [11]. Numbers assigned to stimuli using the category-rating method correspond to the subjective differences between the stimuli.

Magnitude estimation, a ratio measure, is also simple and efficient and has been applied to index construction [12]. On the same continuum for both physical and nonphysical stimuli, category responses are not usually linearly related to magnitude responses. Most often the relation is approximately logarithmic, and such continua have been labeled prothetic. When category responses are linearly related to magnitude responses, the continuum is called metathetic. Magnitude methods usually produce numbers that reflect the subjective ratios between the stimuli, a desirable mathematical property for an index of social consensus [10].

Table 1. Fifty Case Descriptions of Function Status Rated by Health Leaders and Students

Item no.	Age	Mobility	Physical activity	Social activity	Symptom/problem complex
1	40–64	Traveled freely	Walked freely	Performed major and other activities	Had skin defect of face or hands, such as scar, acne, discoloration, or warts
2	≥65	Traveled freely	Walked freely	Performed major and other activities	Had fever, chills, and aching all over
3	<6	Traveled freely	Walked freely	Performed major and other activities	Took medication or followed restrictive diet for preventive purposes
4	18–39	Traveled freely	Walked freely	Performed major and other activities	Breathed unpleasant air in urban area
5	<6	Traveled freely	Walked freely	Performed major but limited in other activities	Had pain in abdomen or side
6	6–17	Traveled with difficulty	Walked freely	Performed major but limited in other activities	Had visual disturbance or impairment, such as nearsightedness or blindness
7	≥65	Traveled freely	Walked freely	Performed major and other activities	Had pain in chest
8	40–64	Traveled freely	Walked freely	Performed major and other activities	Had visual disturbance or impairment, such as nearsightedness or blindness
9	40–64	Traveled freely	Walked freely	Performed major but limited in other activities	Had episodes of feeling hot, and nervous or trembly
10	40–64	Traveled with difficulty	Walked with limitations	Performed major but limited in other activities	Was severely overweight for age and height
11	18–39	Traveled freely	Walked freely	Performed major activity with limitations	Had pain in abdomen or side
12	<6	Traveled freely	Walked freely	Performed major activity with limitations	Had nausea, vomiting, or diarrhea
13	<6	Traveled with difficulty	Walked freely	Performed major activity with limitations	Had visual disturbance or impairment, such as nearsightedness or blindness
14	40–64	Traveled with difficulty	Moved independently in wheelchair	Performed major activity with limitations	Had impairment of one foot or leg, such as fracture, burn, cut, deformity, or paralysis
15	6–17	Traveled with difficulty	Walked with limitations	Performed major activity with limitations	Had pain in back or hips
16	40–64	Traveled freely	Walked freely	Did not perform major but performed self-care activities	Had fever or chills with vomiting or diarrhea
17	≥65	In house	Walked with limitations	Did not perform major but performed self-care activities	Had fever, chills, and aching all over

Table 1 (cont.)

Item no.	Age	Mobility	Physical activity	Social activity	Symptom/problem complex
18	⩾65	In house	Moved independently in wheelchair	Did not perform major but performed self-care activities	Had one foot or leg missing
19	⩾65	Traveled freely	Walked freely	Did not perform major but performed self-care activities	Had nausea, vomiting, or diarrhea
20	⩾65	In house	Moved independently in wheelchair	Did not perform major but performed self-care activities	Had impairment of one arm and leg, such as fracture, burn, cut, deformity, or paralysis
21	6–17	In hospital	Walked freely	Did not perform major but performed self-care activities	Had cough, wheezing, or shortness of breath
22	18–39	Traveled with difficulty	Walked with limitations	Did not perform major but performed self-care activities	Had cough, wheezing, or shortness of breath
23	⩾65	Traveled with difficulty	Walked with limitations	Did not perform major but performed self-care activities	Had impairment of back or spine, such as deformity or weakness
24	⩾65	In hospital	Moved independently in wheelchair	Did not perform major but performed self-care activities	Had impairment of one arm and leg, such as fracture, burn, cut, deformity, or paralysis
25	⩾65	Traveled with difficulty	Moved independently in wheelchair	Did not perform major but performed self-care activities	Had pain in back or hips
26	6–17	Traveled with difficulty	Moved independently in wheelchair	Did not perform major but performed self-care activities	Had impairment of one foot or leg, such as fracture, burn, cut, deformity, or paralysis
27	⩾65	In hospital	In bed or chair	Did not perform major but performed self-care activities	Had pain in abdomen or side
28	40–64	In house	Walked freely	Did not perform major but performed self-care activities	Had one hand or arm missing
29	40–64	In house	Walked freely	Did not perform major but performed self-care activities	Had burning or itching rash on large areas of face, body, or extremities
30	40–64	In house	In bed or chair	Did not perform major but performed self-care activities	Had episodes of feeling hot, and nervous or trembly
31	40–64	In hospital	Walked with limitations	Did not perform major but performed self-care activities	Had burn over large areas of face, body, or extremities
32	⩾65	In house	Moved independently in wheelchair	Required assistance with self-care activities	Had two legs missing

33	40–64	In house	In bed or chair	Required assistance with self-care activities	Had genital pain, bleeding, or discharge
34	40–64	In hospital	Walked freely	Required assistance with self-care activities	Had visual disturbance or impairment, such as nearsightedness or blindness
35	6–17	In house	Walked freely	Required assistance with self-care activities	Had burning or itching rash on large areas of face, body, or extremities
36	18–39	In house	Walked with limitations	Required assistance with self-care activities	Had pain, numbness, or discomfort of feet or legs
37	6–17	In house	Walked with limitations	Required assistance with self-care activities	Had general weakness, fatigue, and weight loss or growth failure
38	<6	In hospital	In bed or chair	Required assistance with self-care activities	Had loss of consciousness such as convulsion, coma, or concussion
39	6–17	In special unit	In bed or chair	Required assistance with self-care activities	Had pain in abdomen or side
40	18–39	In hospital	Walked with limitations	Required assistance with self-care activities	Had cough, wheezing, or shortness of breath
41	6–17	In hospital	Walked with limitations	Required assistance with self-care activities	Had pain in back or hips
42	40–64	In hospital	Walked with limitations	Required assistance with self-care activities	Had impairment of one foot or leg, such as fracture, burn, cut, deformity, or paralysis
43	40–64	Traveled with difficulty	Walked freely	Did not perform major but performed self-care activities	Had visual disturbance or impairment, such as nearsightedness or blindness
44	40–64	In hospital	Moved independently in wheelchair	Required assistance with self-care activities	Had impairment of two legs, such as fracture, burn, cut, deformity or paralysis
45	6–17	In hospital	In bed or chair	Required assistance with self-care activities	Had loss of consciousness such as convulsion, coma, or concussion
46	18–39	In hospital	In bed or chair	Required assistance with self-care activities	Had episodes of feeling hot, and nervous or trembly
47	40–64	In special unit	In bed or chair	Required assistance with self-care activities	Had one hand or arm missing
48	≥65	In special unit	In bed or chair	Required assistance with self-care activities	Had burning or itching rash on large areas of face, body, or extremities
49	<6	In special unit	In bed or chair	Required assistance with self-care activities	Had headache, dizziness, or ringing in ears
50	6–17	In special unit	In bed or chair	Required assistance with self-care activities	Had loss of consciousness such as convulsion, coma, or concussion

Equivalence, an adaptation of the method of adjustment or equivalent stimuli to utility analysis, was devised to represent the implicit trade-offs in health resource allocation and to quantify the comparisons on a ratio continuum of number of persons affected by a health program [13,14]. In this type of procedure subjects adjust variable comparison stimuli until they are subjectively equal to a standard stimulus on a defined continuum. When the value and number of persons in a standard group have been specified, the subjects can express their preferences by adjusting the numbers of persons in each condition described to the point of subjective equality (PSE), or indifference. Assuming that days in all lives can be considered of equal importance, the value assigned to each case description can be derived from the equality judgments.

The relation between values assigned by the category, magnitude, and equivalence methods should give some indication of the convergent validity of the procedure. If the easily administered category and magnitude methods give the same results, then a choice between the two methods must rest on criteria other than internal validity. Equivalence provided a criterion procedure by exposing the implicit values that must be made explicit for a social-welfare function [15]. Thus the category and magnitude procedures could be tested for consistency with a social-choice interpretation of the index.

Method

Two types of judges were interviewed: graduate students and health leaders. Individually tested students, from the Graduate School of Public Administration at New York University, were not medically trained but had received health-related course work and were employed mostly in the health field. Group-tested students, from the Columbia University School of Public Health, all had training in the health field, including nutrition, nursing, and medicine.

Individually tested leaders were members of the New York State Health Planning Commission and Advisory Council appointed by the governor. The Commission included the commissioners and executive personnel of the ten major health-related departments of state government—health, mental hygiene, labor, insurance, social services, education, environmental conservation, agriculture, local government, and planning services. The Advisory Council, which included both health professionals and nonmedical consumer representatives, exercises value judgments concerning the health problems of New York State; its members participated because of their interest and experience in health decision making. Group-tested leaders were predominantly consumer members of the NY-Penn Health Planning Council in Binghamton, N.Y., an agency that serves in the same capacity on an areawide level as the state Advisory Council.

The 50 case descriptions in Table 1 were typed on separate sheets in the following form:

40–64
Walked with limitations

In hospital
Did not perform major activity but performed self-care activities
Had burn over large areas of face, body, or extremities

These were inserted randomly into a testing booklet for each judge. Ten warm-up items, given in a single random order for each scaling method, familiarized the judges with the entire range of the scale and with the assignment of numbers proportional to subjective feelings. Each item was written like a diary of observations made at the end of a single day, with each description referring to a different day. This focus on "days of dysfunction" was made in order to eliminate the prognostic dimension of health status from the judges' responses.

Instructions and Scoring

Judges for all scoring methods were instructed to ignore adjacent days and to confine their judgments, as far as possible, to the single day of dysfunction described in each item, to evaluate the severity of the symptom or problem from its effect on the other indicators of function (since the disturbances could be caused by a variety of diseases), and to assume that the person described was receiving optimal treatment and performed as well as possible on that day. They were also instructed *not* to turn back once they had evaluated a case.

Items scored by the category method contained a numerical rating scale on each page, described as being constructed of 11 equal intervals. Instructions for this procedure were as follows:

> Evaluate the desirability of each day by circling a number from 1 to 11 which shows how desirable each day seems to you. Each number represents an equal step on a scale of desirability such that 5 is one step more desirable than 4, 11 is one step more desirable than 10, and so forth. The label "most desirable" is above category 11 and represents a day in the life of a person who was as healthy as possible on that day, i.e., performed his major and other activities, had no discernible symptoms, and walked and traveled about freely. The label "least desirable" is below category 1 and represents a person who died during the day. All items fall between these two extremes, and you may use all 11 categories as you see fit.

For magnitude estimation, a standard item representing the upper extreme of the scale was given a score of 1000. The instructions were:

> Evaluate the desirability of each day by writing in the score box a number which reflects how preferable each day seems to you. This standard item describes a day which has been given a score of 1000. It is a day in the life of a person who was as healthy as possible on that day. Every other day should be scored in relation to this standard description. For example, if the item seems half as desirable as the standard, then write in a score of 500. If the day appears a tenth as preferable as the standard, then write in a score of 100. You may use any whole number or fraction that is greater than zero and equal to or less than 1000.

For equivalence, the standard case description was the same but the scoring method was different. Judges were told they were going to play a decision-

Table 2. Number and Classification of Judges in 14 Experimental Groups

Method	Order of presentation	Students			Leaders		
		Group code	Group no.	*N*	Group code	Group no.	*N*
Individual							
Category	First	C1S	1	46	C1L	3	23
	Second	C2S	2	30	C2L	4	23
Magnitude	First	M1S	7	47	M1L	9	45
	Second	M2S	8	32	M2L	10	23
Equivalence	Second	E2S	13	28	E2L	14	18
Group:							
Category	...	CGS	5	26	CGL	6	40
Magnitude	...	MGS	11	22	MGL	12	60
				231			232

making game with assumptions they might consider unrealistic. The equivalence instructions were:

> Suppose there are two groups of people, both of which will die immediately if not helped. You have the resources to keep one and only one of these groups alive for one more year, after which they will also die. The first group contains 100 people in a state of maximum health (standard). I want you to make a decision concerning the number of people in the second group. Persons in the second group are in a state of health lower than the standard (items in the booklet). With each item in this booklet, ask yourself this question: "How many people in this state of health do I consider equivalent to the 100 people of the same age in the standard group?" Start with 100 and increase this number to the point at which you are not able to decide between the standard and comparison groups. You may use any number equal to or greater than 100.

Individually tested students and leaders were randomly assigned to two of the three methods and to different orders of method presentation—the students to Groups 1, 2, 7, 8, and 13 and the leaders to Groups 3, 4, 9, 10, and 14, as shown in Table 2. Because some judges had difficulty performing the equivalence procedure without previous experience with category or magnitude, equivalence was presented only as a second method. Group-tested judges did not perform the equivalence procedure but were randomly assigned to either the category or the magnitude method, the students to Groups 5 or 11 and the leaders to Groups 6 or 12. To evaluate the internal consistency of the measurement methods, the 50 case descriptions were randomly repeated in sets of 5 each at the end of all booklets for the group-tested judges (Groups 5, 6, 11, and 12).

Results and Analysis

The evaluation of one item by one judge constitutes a single datum in a frequency matrix partitioned by (1) judge type, (2) scaling method, (3) order of method presentation, and (4) interview situation. Assuming that the judges assigned values according to the instructions, category responses were normalized by assigning the values computed for the midpoints of the 11 intervals on a 0–1 scale; magnitude responses were divided by 1000, the number assigned to the complete-well-being anchor at the upper extreme. The equivalence scale values were derived from the basic equation

$$E_s N_s = E_j N_j$$

where $N_s = 100$, the number of persons in the standard case describing optimal function
$E_s = 1.0$, the value assigned to optimal function
$E_j =$ preference value for comparison case j
$N_j =$ number of persons in case j assigned by the judge

From this basic equation, the equivalence value E_j assigned to each case on a 0–1 scale can be derived from $100/N_j$, the point of subjective equality between the standard case and the comparison cases.

Table 3 (overleaf) contains the means and standard deviations of the preferences for the 50 items made by the 14 experimental groups. Item means ranged across methods from .12 to .91; most fell near the center of the scale. The standard deviations did not follow any pattern according to item position on the preference continuum for any of the three methods. In general, standard deviations for equivalence responses were higher than for responses obtained using the category or magnitude methods. The sample standard deviations of judges' responses to a single item varied from .07 to .33; the standard deviation of the item means among different groups was usually less than .03.

The multimethod-multigroup intercorrelation matrix revealed that most of the correlation coefficients were above .95 and all were above .90. Such results, while impressive, do not adequately distinguish the groups. So a multivariate discriminant analysis of variance was used to compute the matrix of F statistics (Table 4, p. 240) in order to test the equality of the vector of item means between pairs of the 14 groups [16]. Comparisons among the randomized groups provide direct tests of the effects of order and of method and their interactions. Although tests between students and leaders and between individual and group interviews were not randomized, reasonable inferences are possible when comparisons are made where all other factors are held constant.

Table 4 groups the array of F ratios by the three methods. For 78 simultaneously observed Fs, the .05 significance level gives a critical value of 1.86. (For multiple comparisons, the significance level that gives an appropriate overall error rate of α is approximated by α/n, where n is the number of simultaneous comparisons. For $\alpha = .05$ and 78 comparisons, the level of significance is .0064. The critical value of F was computed using an approximation to the inverse function for the F distribution [17,18].) The monomethod-multigroup category triangle, the upper left corner of the matrix (Groups 1–6), reveals no significant

Table 3. Means and Standard Deviations of Preferences for 50 Case Descriptions of Function Scaled by 14 Experimental Groups

Item no.	Statistic	Experimental group number and code													
		1 C1S	2 C2S	3 C1L	4 C2L	5 CGS	6 CGL	7 M1S	8 M2S	9 M1L	10 M2L	11 MGS	12 MGL	13 E2S	14 E2L
1	Mean	.78	.78	.79	.74	.74	.86	.81	.85	.87	.81	.82	.87	.71	.68
	S.D.	.18	.11	.11	.16	.17	.13	.20	.14	.09	.10	.13	.13	.33	.29
2	Mean	.65	.66	.57	.62	.57	.72	.72	.75	.72	.69	.66	.74	.53	.52
	S.D.	.19	.20	.19	.17	.17	.13	.18	.14	.15	.15	.18	.18	.29	.31
3	Mean	.77	.79	.78	.82	.76	.83	.87	.84	.89	.76	.79	.84	.72	.78
	S.D.	.17	.14	.14	.10	.19	.23	.12	.18	.08	.17	.18	.13	.31	.26
4	Mean	.85	.86	.76	.84	.80	.91	.89	.88	.91	.84	.80	.90	.87	.83
	S.D.	.12	.09	.18	.09	.15	.07	.17	.15	.09	.12	.16	.08	.20	.20
5	Mean	.59	.60	.59	.59	.53	.62	.66	.66	.68	.63	.57	.71	.58	.55
	S.D.	.20	.17	.15	.17	.16	.28	.19	.19	.16	.17	.19	.17	.29	.29
6	Mean	.52	.50	.47	.44	.43	.43	.58	.54	.54	.50	.49	.55	.42	.45
	S.D.	.17	.17	.15	.15	.18	.25	.19	.17	.18	.20	.19	.22	.28	.27
7	Mean	.65	.68	.57	.62	.63	.71	.69	.73	.73	.71	.60	.66	.53	.53
	S.D.	.21	.18	.21	.14	.18	.21	.21	.18	.13	.15	.22	.22	.31	.32
8	Mean	.66	.67	.65	.62	.58	.66	.75	.74	.71	.72	.69	.68	.53	.51
	S.D.	.18	.15	.16	.16	.23	.25	.17	.16	.18	.14	.20	.21	.28	.30
9	Mean	.67	.58	.56	.60	.63	.66	.69	.70	.70	.62	.65	.70	.52	.56
	S.D.	.15	.14	.17	.16	.16	.20	.17	.15	.14	.12	.18	.18	.29	.28
10	Mean	.52	.44	.48	.48	.42	.46	.55	.56	.55	.52	.45	.51	.37	.42
	S.D.	.20	.15	.18	.16	.17	.28	.19	.17	.19	.19	.20	.22	.26	.28
11	Mean	.63	.61	.58	.61	.60	.70	.70	.67	.69	.63	.59	.69	.60	.50
	S.D.	.18	.16	.17	.14	.15	.17	.16	.19	.13	.17	.16	.20	.30	.28
12	Mean	.61	.58	.53	.59	.56	.59	.69	.68	.67	.62	.61	.68	.56	.50
	S.D.	.19	.21	.20	.19	.21	.26	.18	.20	.17	.17	.19	.20	.32	.27
13	Mean	.48	.44	.48	.41	.41	.44	.53	.52	.47	.49	.46	.50	.45	.40
	S.D.	.20	.16	.20	.13	.20	.25	.24	.21	.15	.19	.19	.21	.29	.25
14	Mean	.43	.42	.43	.44	.35	.44	.54	.49	.48	.48	.46	.49	.37	.34
	S.D.	.17	.16	.12	.12	.11	.21	.14	.18	.17	.14	.17	.21	.25	.26
15	Mean	.49	.45	.41	.42	.36	.45	.53	.48	.50	.46	.45	.47	.39	.39
	S.D.	.16	.16	.16	.17	.14	.24	.18	.21	.19	.16	.17	.20	.25	.26
16	Mean	.58	.54	.51	.54	.57	.55	.66	.66	.61	.59	.59	.62	.46	.38
	S.D.	.19	.16	.15	.16	.15	.19	.16	.16	.18	.12	.18	.20	.31	.25
17	Mean	.46	.38	.46	.44	.40	.42	.53	.52	.51	.45	.48	.52	.34	.31
	S.D.	.16	.16	.19	.14	.17	.23	.20	.14	.16	.15	.16	.21	.30	.24
18	Mean	.43	.40	.46	.40	.36	.49	.55	.47	.45	.51	.46	.48	.29	.26
	S.D.	.17	.18	.13	.12	.17	.20	.17	.14	.18	.17	.18	.20	.25	.19
19	Mean	.60	.57	.56	.59	.60	.61	.67	.68	.64	.58	.66	.66	.43	.41
	S.D.	.17	.17	.17	.17	.17	.19	.15	.14	.17	.16	.18	.20	.31	.29
20	Mean	.42	.39	.42	.41	.36	.51	.52	.49	.45	.49	.46	.51	.32	.29
	S.D.	.16	.16	.15	.14	.14	.20	.15	.14	.17	.17	.17	.20	.25	.22
21	Mean	.43	.34	.37	.38	.32	.41	.49	.44	.44	.40	.39	.49	.36	.31
	S.D.	.17	.15	.12	.16	.14	.25	.20	.19	.16	.10	.16	.20	.27	.20
22	Mean	.47	.40	.40	.43	.37	.43	.51	.49	.45	.42	.43	.44	.40	.27
	S.D.	.16	.15	.12	.14	.15	.25	.19	.18	.17	.14	.17	.18	.28	.20
23	Mean	.45	.40	.43	.41	.37	.46	.47	.51	.44	.45	.44	.47	.33	.30
	S.D.	.17	.18	.16	.11	.15	.18	.19	.17	.18	.16	.18	.21	.24	.24
24	Mean	.38	.35	.36	.36	.35	.41	.44	.41	.41	.42	.41	.45	.28	.25
	S.D.	.18	.13	.14	.12	.15	.21	.19	.12	.17	.13	.18	.20	.21	.21
25	Mean	.44	.42	.41	.39	.40	.44	.53	.52	.45	.49	.47	.52	.31	.29
	S.D.	.17	.17	.14	.13	.14	.19	.14	.17	.17	.13	.17	.23	.21	.23

Table 3 (cont.)

Item no.	Statistic	Experimental group number and code													
		1 C1S	2 C2S	3 C1L	4 C2L	5 CGS	6 CGL	7 M1S	8 M2S	9 M1L	10 M2L	11 MGS	12 MGL	13 E2S	14 E2L
26	Mean	.45	.37	.44	.40	.35	.40	.49	.46	.46	.43	.44	.45	.36	.28
	S.D.	.17	.17	.14	.14	.15	.20	.18	.16	.16	.15	.19	.21	.25	.24
27	Mean	.33	.29	.36	.36	.31	.36	.43	.46	.42	.42	.37	.41	.24	.24
	S.D.	.10	.14	.14	.13	.12	.19	.19	.16	.18	.16	.17	.21	.21	.19
28	Mean	.44	.41	.46	.42	.39	.48	.54	.51	.49	.49	.47	.49	.33	.27
	S.D.	.17	.14	.13	.14	.15	.21	.19	.17	.18	.16	.17	.20	.25	.21
29	Mean	.51	.46	.45	.45	.42	.46	.55	.54	.55	.48	.53	.57	.38	.30
	S.D.	.18	.14	.16	.15	.19	.20	.20	.14	.18	.14	.20	.21	.30	.23
30	Mean	.46	.38	.46	.44	.42	.45	.52	.50	.50	.47	.47	.54	.37	.28
	S.D.	.17	.13	.16	.16	.17	.22	.19	.17	.18	.14	.19	.22	.27	.23
31	Mean	.34	.33	.32	.32	.26	.28	.40	.38	.38	.36	.30	.36	.28	.25
	S.D.	.14	.16	.18	.13	.14	.22	.18	.16	.20	.12	.19	.20	.25	.19
32	Mean	.30	.30	.35	.33	.30	.35	.40	.36	.36	.48	.40	.36	.23	.23
	S.D.	.17	.15	.14	.15	.19	.21	.19	.17	.19	.17	.19	.20	.18	.17
33	Mean	.35	.32	.31	.38	.31	.37	.40	.44	.42	.40	.37	.43	.34	.27
	S.D.	.14	.13	.13	.14	.10	.21	.18	.17	.18	.16	.17	.20	.28	.22
34	Mean	.41	.33	.36	.33	.34	.36	.38	.45	.42	.39	.44	.43	.31	.27
	S.D.	.17	.13	.13	.12	.18	.21	.18	.15	.18	.14	.18	.20	.25	.20
35	Mean	.44	.43	.43	.41	.40	.43	.52	.51	.51	.47	.51	.53	.41	.35
	S.D.	.18	.18	.15	.15	.18	.18	.22	.19	.19	.17	.20	.22	.29	.27
36	Mean	.43	.32	.36	.40	.31	.39	.46	.45	.44	.43	.37	.45	.34	.26
	S.D.	.15	.15	.11	.15	.13	.23	.19	.19	.17	.13	.17	.20	.25	.18
37	Mean	.35	.32	.31	.36	.26	.31	.38	.39	.37	.38	.30	.36	.34	.26
	S.D.	.15	.17	.12	.13	.11	.24	.20	.20	.18	.18	.15	.17	.26	.19
38	Mean	.19	.19	.21	.23	.14	.19	.20	.24	.23	.23	.22	.25	.26	.26
	S.D.	.11	.16	.12	.17	.08	.22	.15	.17	.17	.12	.17	.16	.22	.15
39	Mean	.29	.25	.28	.27	.30	.29	.38	.29	.34	.32	.30	.34	.36	.28
	S.D.	.14	.16	.13	.16	.19	.25	.20	.18	.17	.15	.19	.21	.27	.24
40	Mean	.34	.27	.33	.34	.28	.32	.38	.39	.34	.36	.34	.38	.29	.23
	S.D.	.16	.11	.15	.16	.13	.21	.20	.15	.16	.10	.17	.19	.24	.17
41	Mean	.36	.28	.32	.33	.29	.33	.37	.36	.38	.39	.32	.41	.32	.33
	S.D.	.14	.16	.12	.18	.11	.21	.18	.17	.19	.12	.18	.19	.25	.17
42	Mean	.36	.29	.34	.32	.30	.32	.41	.35	.35	.35	.38	.40	.29	.29
	S.D.	.16	.14	.14	.14	.12	.21	.20	.16	.15	.11	.20	.20	.26	.18
43	Mean	.53	.42	.48	.43	.45	.43	.54	.57	.52	.51	.52	.50	.37	.29
	S.D.	.18	.15	.13	.14	.18	.23	.19	.15	.14	.15	.17	.21	.26	.21
44	Mean	.34	.31	.27	.29	.27	.27	.36	.34	.33	.35	.36	.41	.25	.22
	S.D.	.16	.17	.09	.12	.14	.17	.15	.14	.16	.12	.16	.17	.18	.17
45	Mean	.21	.21	.19	.23	.19	.21	.24	.27	.27	.26	.22	.25	.26	.19
	S.D.	.10	.16	.11	.17	.11	.22	.17	.18	.16	.14	.13	.17	.17	.14
46	Mean	.36	.28	.31	.33	.31	.32	.37	.39	.41	.40	.35	.41	.34	.31
	S.D.	.15	.11	.12	.12	.14	.24	.21	.17	.19	.13	.21	.20	.25	.24
47	Mean	.27	.25	.29	.28	.23	.25	.30	.29	.32	.30	.29	.30	.27	.23
	S.D.	.15	.14	.14	.14	.12	.23	.18	.16	.18	.14	.17	.18	.22	.16
48	Mean	.30	.22	.31	.27	.22	.23	.31	.29	.32	.29	.28	.32	.25	.23
	S.D.	.16	.12	.18	.16	.12	.21	.19	.17	.19	.16	.19	.21	.22	.17
49	Mean	.27	.21	.25	.27	.26	.20	.28	.29	.29	.30	.26	.28	.31	.25
	S.D.	.14	.18	.15	.17	.15	.20	.20	.18	.18	.14	.17	.16	.24	.18
50	Mean	.20	.16	.19	.21	.16	.12	.20	.21	.23	.23	.18	.20	.24	.19
	S.D.	.11	.14	.11	.16	.08	.11	.17	.17	.17	.14	.14	.16	.17	.13

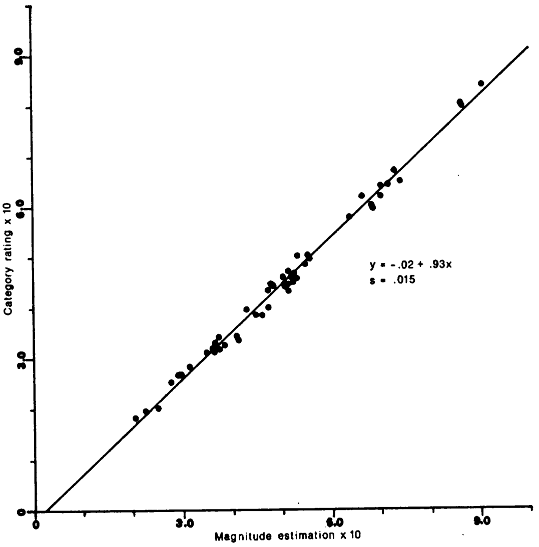
Table 4. Multigroup-Multimethod Matrix of F Values Testing Equality of Vector of 50 Item Means between Pairs of 14 Experimental Groups

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
	C1S	C2S	C1L	C2L	CGS	CGL	M1S	M2S	M1L	M2L	MGS	MGL	E2S	E2L
1 Category rating First/Students														
2 Category rating Second/Students	.83													
3 Category rating First/Leaders	1.35	1.09												
4 Category rating Second/Leaders	1.05	.71	.72											
5 Category rating Group/Students	1.03	.87	1.13	1.01										
6 Category rating Group/Leaders	1.15	.92	1.10	1.09	.96									
7 Magnitude est. First/Students	1.39	.94	1.45	.95	1.51	1.62								
8 Magnitude est. Second/Students	1.12	.88	1.23	.93	1.30	1.09	1.08							
9 Magnitude est. First/Leaders	1.02	.85	1.34	.75	1.32	1.39	1.26	.89						
10 Magnitude est. Second/Leaders	1.05	.68	.98	.86	1.00	1.21	.92	.56	.75					
11 Magnitude est. Group/Students	1.41	.99	1.37	1.09	1.32	1.36	1.17	.99	1.12	.79				
12 Magnitude est. Group/Leaders	1.57	1.18	1.66	1.23	1.68	1.49	1.61	1.07	1.07	.87	.86			
13 Equivalence Second/Students	1.45	1.04	1.50	.90	1.09	1.78	2.12*	1.72	1.62	1.27	1.43	1.93*		
14 Equivalence Second/Leaders	1.43	.96	1.28	.93	1.04	1.59	1.93*	1.89*	1.38	1.42	1.59	1.86*	.91	

* $p < .05$ ($F_{.0064,\,50,\,400} = 1.86$).

differences between students and leaders, between the orders of method presentation, between group and individual interviews, or between interactions of these three factors. Similar results arise from the multigroup-monomethod triangle of magnitude responses (Groups 7–12). Equivalence judgments were not performed as a first method or with group interviews, but the individual interviews (Groups 13 and 14) again show no significant difference between students and leaders.

In the block of 36 multigroup-heteromethod comparisons between the category and magnitude methods (Groups 1–6 vs. 7–12), no significant differences were detected either in direct comparisons or with interactions among the treatment factors, further substantiating the results obtained above. The curve relating the category and magnitude scales, as shown in the accompanying figure, is clearly a straight line. In addition, no significant differences were detected for the block of 12 multigroup-heteromethod comparisons between the category and equivalence methods (Groups 1–6 vs. 13–14). Overall, the category method produced results similar to those obtained by magnitude and equivalence judgments. The F ratios were smallest for category rating as a second method, indicating the beneficial effect of previous scaling experience on the reliability of the subjects' responses.



Mean magnitude and category scores for 50 items scaled by student and leader experimental groups.

Multigroup-heteromethod comparisons between the magnitude and equivalence methods (Groups 7–12 vs. 13–14) revealed 5 significant differences in 12 comparisons. Although the randomized comparisons of these two methods (E2S vs. M2S and E2L vs. M2L) were not significant, equivalence judgments were apparently less resistant to interaction effects, at least with respect to magnitude estimation.

Internal consistency in scoring the items was tested by repeating certain items in the booklets for all judges who participated in the group interviews. Judges reported that the complexity of the stimuli prevented their remembering initial scores when making repeat judgments. Pearson's correlation coefficients between initial and repeat judgments for each group were as follows:

Category - Group - Students:	.83
Category - Group - Leaders:	.79
Magnitude - Group - Students:	.83
Magnitude - Group - Leaders:	.74

Student's *t* tests for related observations across all the judges were not significantly different ($p < .05$).

Agreement among the judges on the relative position and distance of the items on the scale was tested by computing Pearson's *r* for each judge against the pooled group item means for each method. With judges who did two methods counted twice, the mean *r* for each scale was as follows: category rating ($N = 122$): .77; magnitude estimation ($N = 147$): .79; equivalence ($N = 46$): .60; category rating ($N = 66$): .75; magnitude estimation ($N = 82$): .75.

To use a health index for resource allocation and evaluative research models, an equal-interval scale becomes essential. Although this property is difficult to demonstrate mathematically, one highly recommended procedure is the functional measurement test [11]. In this test, four items are selected so that two pairs differ only in their symptom/problem complex, while the two items with the same complex differ in function-status descriptors. Since all other factors are held constant in the pairs of items, and assuming no interaction, their differences on an equal-interval scale should be zero. Marginal means were computed across all the groups by scaling method for items 25, 26, 41, 42, 28, 29, 47, and 48. Means for items 41 and 42 were subtracted from the means for items 25 and 26. Means for items 47 and 48 were subtracted from the means for items 28 and 29. The differences between item-pair differences were then calculated. These differences between the item pairs—ranging from .018 to .061—were within the sampling error for item means for all methods. The parallel structure revealed by this test is accepted evidence for the interval nature of the obtained scale.

Discussion

The absence of significant differences in the values assigned to the items between directly comparable groups provides evidence for the convergent validity of the measurement methods and results. Although the interjudge variation in the item responses may be obscuring some differences among the methods and groups that would be detected with larger numbers of judges, this convergence establishes the feasibility of measuring the mean level of well-being that social groups associate with descriptions of function status. The similarity of equivalence to category and most magnitude groups indicates the consistency of these methods with a procedure implying social choice.

Further evidence for validity would require that all the methods be compared with other choice procedures, such as paired comparisons or the Von Neumann-Morgenstern expected-utility model, and with data from actual social choices. The method of equivalent stimuli represents a step in that direction.

As illustrated in the figure, the relation between the category and magnitude scales is clearly linear. This result suggests that the preference continuum for conditions of function status may be metathetic. This conclusion is tentative, however, since the magnitude estimation scale was adapted to the function status continuum by anchoring it at the upper extreme with the description of a perfectly well day. Since this is not the usual practice with physical stimuli, where the prothetic/metathetic distinction was established, the standard at the upper scale extreme could have restricted the judges and effectively converted the magnitude scale to a category rating scale from 0 to 1000. On the other hand, most prothetic continua such as loudness or brightness do not have a conceptual limit such as a well day. Only further study with unconstrained judges can determine whether this adaptation of magnitude estimation had a significant effect.

The reliability of the measurement methods is indicated by the stability of the values across different orders of method presentation, individual and group interviews, student and leader judges, and interactions of all these variables. The agreement between the graduate students and the health leaders is understandable in view of the similar backgrounds and career interests of the two groups.

The correlation of approximately .81 between the judges' separate responses on the same items can be used to estimate an overall reliability coefficient on single items of .90, which is equivalent to the correlation of an average single response with the long-run "true" mean [19]. Combining several items for multiple judges will make it feasible on household surveys to establish function-level means at a 99 percent confidence interval of .01 on a 0–1 scale, a reasonable range for health index values.

The disparity between several of the magnitude and equivalence groups, however, indicates that these two methods are more sensitive to changes in experimental context than the category procedure, which showed no differences among any of the groups and methods. In the present context, therefore, category scaling appears to be more stable across modifications in procedure.

The methods themselves did not provide a means of testing whether preference values can reasonably be represented as a unidimensional continuum. But the high degree of agreement among the judges concerning the rank order and scale separation of the items supports the hypothesis that unidimensional judgments were obtained. If judges had responded to an unknown number of different dimensions, the relative values of scale items would have varied from judge to judge and from method to method. On the basis of data at hand, the values evidenced remarkable stability across different judges and scaling methods, supporting the conclusion that unidimensionality was achieved. (Although Guttman scaling and factor analysis are widely used in analyzing the dimensionality of data from response methods, they are not considered appropriate for analyzing data from judgment methods where subjects are instructed to scale stimuli according to a particular attribute, in this case preference [20].)

The functional measurement test of equal intervals established the interval properties of the separate scales. These scales can then be used to weight the distribution of life expectancies among the different function levels to quantify the effectiveness of a health program. This will permit comparisons among alternative health services.

Both the category and magnitude estimation procedures are easily performed and should be feasible for household interview surveys. Although only a small number of items were repeated with different scaling procedures in this methodological study, household surveys would include a large number of items with a single method, so that the means of social preferences for a series of function levels could be computed. The mean function level values could then be used to compute the function and health status of populations and to evaluate and plan health programs [21,22].

The method of equivalent stimuli is too complex for use outside a laboratory-like individual interview. The unrealistic assumptions and the emotive nature of the task confused and offended some judges. It is unlikely, therefore, that equivalence or other choice methods could be adapted for survey purposes, although they will continue to be useful as criterion procedures.

In conclusion, the value measurement strategy developed in this study has demonstrated the feasibility of measuring levels of well-being for case descriptions of function status. Although the values obtained in this study represent the preferences of a prestigious leadership sample, household interview surveys could provide statistically representative social values for a series of operationally defined function levels. These weights could then be used to incorporate a value component into function status so that health status indexes could be computed for nations or for smaller demographic groups.

As derived here, these values also have interval-scale properties that would make them useful for evaluative research or for projecting and comparing the probable benefits of alternative health programs. Before these goals can be achieved, however, further methodological research is needed to develop measurement methods that will more fully satisfy criteria for validity, reliability, and generalizability of social preferences. From these future efforts should come a measurement method that will permit integration of the previously immeasurable value dimension into a verifiable indicator of health status.

Acknowledgments. The authors wish to thank the students and health leaders who performed the measurement methods; Edward H. Van Ness, C. Addison Keeler, and Richard Landis, who arranged the participation of the health leaders; and Martina Folsom, James Thomas, Alfred Tsai, W. Ray Mickey, and W. R. Blischke, who assisted in the data collection and analysis. Computing assistance was obtained from the Health Sciences Computing Facility, UCLA, sponsored by NIH Special Research Resources Grant RR-3.

References

1. Sullivan, D. F. *Conceptual Problems in Developing an Index of Health.* Vital and Health Statistics Series 2 No. 17. Rockville, Md.: National Center for Health Statistics, 1966.
2. Moriyama, I. M. Problems in the Measurement of Health Status. In E. Sheldon and W. Moore (eds.), *Indicators of Social Change*, p. 573. New York: Russell Sage Foundation, 1968.
3. Fanshel, S. and J. W. Bush. A health-status index and its application to health-services outcomes. *Op. Res.* 18:1021 November–December 1970.
4. Bush, J. W., S. Fanshel, and M. M. Chen. Analysis of a tuberculin testing program using a health status index. *Socio-Econ. Planning Sci.* 6:49, 1972.
5. Williamson, J. W., M. Alexander, and G. E. Miller. Priorities in patient-care research and continuing medical education. *JAMA* 204:303 Apr. 22, 1968.
6. Packer, A. H. Applying cost-effectiveness concepts to the community health system. *Op. Res.* 16:227 March–April 1968.
7. Patrick, D. L., J. W. Bush, and M. M. Chen. Toward an operational definition of health. *J. Health & Soc. Behav.* 14:6 March 1973.
8. Campbell, D. T. and D. W. Fiske. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.* 46:81, 1959.
9. Centra, J. A. Validation by the multigroup-multiscale matrix: An adaptation of Campbell and Fiske's convergent and discriminant validation procedure. *Educ. & Psychol. Meas.* 31:675, 1971.
10. Stevens, S. S. Ratio Scales of Opinion. In D. K. Whitla (ed.), *Handbook of Measurement and Assessment in Behavioral Sciences*, p. 171. Reading, Mass.: Addison-Wesley, 1968.
11. Anderson, N. H. Integration theory and attitude change. *Psychol. Rev.* 78:171, 1971.
12. Sellin, T. and M. E. Wolfgang. *The Measurement of Delinquency.* New York: Wiley, 1964.
13. Torgerson, W. S. *Theory and Methods of Scaling*, p. 141. New York: Wiley, 1958.
14. Guilford, J. P. *Psychometric Methods*, p. 24. New York: McGraw-Hill, 1954.
15. Arrow, K. *Social Choice and Individual Values*, p. 112. New York: Wiley, 1963.
16. Winer, B. J. *Statistical Principles in Experimental Design*, 2d ed., p. 232. New York: McGraw-Hill, 1971.
17. Kendall, M. G. and A. Stuart. *The Advanced Theory of Statistics*, Vol. 3, p. 44. New York: Hafner, 1968.
18. Abramowitz, M. and I. Stegun. *Handbook of Mathematical Functions*, p. 947. Applied Mathematics Series No. 55. Washington: National Bureau of Standards, 1964.
19. Nunnally, J. C. *Tests and Measurements: Assessment and Prediction*, p. 177. New York: McGraw-Hill, 1959.
20. Torgerson, W. S. *Theory and Methods of Scaling*, p. 298. New York: Wiley, 1958.
21. Torrance, G. W., W. H. Thomas, and D. L. Sackett. A utility maximization model for evaluation of health care programs. *Health Serv. Res.* 7:118 Summer 1972.
22. Bush, J. W., M. M. Chen, and D. L. Patrick. Social indicators for health based on function status and prognosis. In *Proceedings of the American Statistical Association, Social Statistics Section*, p. 71. Washington, D.C.: The Association, 1972.