COMPARISON OF METHODS FOR MEASURING SOCIAL PREFERENCES FOR A HEALTH STATUS INDEX

James W. Bush*, Robert M. Kaplan**, and Charles C. Berry*

University of California, San Diego*, and San Diego State University**

INTRODUCTION

THE HEALTH STATUS INDEX

In previous publications, a health status index has been described which could be an effective tool in health planning, health program evaluation, and population monitoring [Patrick et al. 1973a, 1973b; Bush et al. 1973; Chen et al. 1973, 1975; Blischke et al. 1975; Chen and Bush 1976; Kaplan et al. 1976].

The index separates two distinct components: Levels of Well-Being, the weights, social values, or utilities that members of society associate with a person's level of functioning at some point in time, and prognoses -- the probabilities of transition to other levels of function and Well-Being on future occasions. Treating these components as analytically distinct allows the quantitative expression of the two variables.

Since the quantities vary independently, joint functions of the two variables are necessary to fully describe health status. Thus, no precise statement of health status can be made for an individual or a group without knowledge of the expected transitions among the function levels over time. We shall, therefore, reserve the term "health" for a composite expression of prognosis and function level as well as Level of Well-Being.

The present report concerns the utility dimension of health. This is the social preference or "Level of Well-Being" for states of function on a continuum from optimum function (1.0) to death (0.0). When these weights have been measured, health status can be expressed precisely as the expected value (product) of the preferences associated with the states of function at a point in time and the probabilities of transition to other states over the remainder of the life expectancy [Kaplan et al. 1976].

Steps from three scales -- Mobility, Physical Activity, and Social Activity -- can be combined into sets called Function Levels.* Any individual can be classified into one of the mutually exclusive and collectively exhaustive Function Levels. Subjective, symptomatic disturbances are incorporated in an independent set of symptom/problem complexes whose presence or absence can be noted in surveys and follow-up studies.

Levels of Well-Being are the weights, social preferences, or measures of relative importance that members of society associate with each of the Function Levels. These preferences may be measured by having consumers rate sets of standardized but realistic case descriptions. The case descriptions consist of the items of information describing a Function Level and a Symptom/Problem Complex, and describe how a person would be classified according to the items in the Index. Thus, unlike weights obtained from arbitrary, disease specific scenarios, the weights obtained can be assigned with little error to all actual persons.

Since utilities are an important component of the Index of Well-Being, accurate, reliable ratings on an interval or ratio scale of measurement are highly desirable. This study compares results obtained via magnitude estimation, a method purported to yield ratio scales, with data obtained by a simpler, more widely accepted method known as category rating.

PREFERENCE MEASUREMENT

In a number of his publications, S.S. Stevens refers to two classes of psychological continua: prothetic describes intensity dimensions such as light or sound, and metathetic describes qualitative dimensions, such as pitch or visual position. The functional form of the responses among the scaling procedures determines the type of continuum.

Category rating is a simple partition method in which subjects are requested to assign each stimulus to a set of numbered categories representing equal intervals. This method, exemplified by the familiar 10-point rating scale, is efficient, easy to use, and applicable in a large number of laboratory and survey settings. Stevens [1966, 1971, 1974] questioned the assumption that the subjective impressions of a stimulus can be discriminated equally at each level of the scale. With Galanter [1957] he claimed that the category method is biased because subjects attempt to use each category equally often -- spreading out the ratings when the stimuli are actually close together, and pushing them together when the true values are far apart.

In a long series of studies, the same authors [1957] purportedly demonstrated that the results of magnitude estimation accurately represent sensory and nonsensory perceptions. With this procedure, a subject is given a standard stimulus and asked to provide a subjective ratio by assigning numbers to other stimuli "in proportion to" the number assigned to the standard case. Except in rare cases, the mean category ratings are linearly related to the logarithms of the arithmetic or geometric mean magnitude estimation judgments.

The present analysis extends a previous study [Patrick et al. 1973b] which described a linear relationship between magnitude estimation and category rating. That study could be criticized, however, because a standard (Well-Day) for magnitude estimation was assigned the value 1000 to represent the top extreme of the scale. The bounding of the scale, which is not standard in magnitude estimation, might have forced the linear relationship because it effectively made the procedure a form of category rating. The present study examines the relationship between category scaling and an unbounded form of magnitude estimation.

* See Appendix I.

METHOD

SUBJECTS AND CASE DESCRIPTIONS

The subjects were 65 volunteers from introductory psychology courses at San Diego State University with roughly equal proportions of males and females.

The items or case descriptions were drawn from a sample frame which includes all possible combinations of Function Levels and Symptom/Problem Complexes. Since age is necessary to provide a meaningful case description but contributes little to the variance of the ratings [Chen et al. 1973], one of four age groups was also identified with each item.

Thirty items were chosen to represent the full range of dysfunctions imposed on all types of patients by multiple symptoms and problems, including near well states. Each step in the scales of Mobility (MOB), Physical Activity (PAC), and Social Activity (SAC) was included at least once in the set of case descriptions. The first five items included a description of a completely well person and a person in a comatose state. These items familiarized the subjects with scale extremes. In sum, each item is a combination of an age group, one step from each of the three scales, and one symptom/problem complex (CPX), as follows:

School age (6-17),	(AGE)
Used car, bus or train as usual for age,	(MOB)
Walked with physical limitations,	(PAC)
Limited in amount or kind of school work,	(SAC)
Had pain, bleeding, itching or discharge from sexual organs.	(CPX)

The stimuli were presented as single pages in thirty item booklets. The content of the items within each booklet was identical and the order of the first five (warm-up) items was constant. The study items, however, were in a computer generated random order. Half the subjects were assigned to do category first, and the other half to do magnitude first, using different booklets for the two procedures. The subjects were run in groups of three to five students. Detailed instructions are available from the authors.

DATA CLEANING

A set of rules was created to eliminate judges who had apparently not paid close attention or did not understand the instructions. The rules eliminated subjects who rated two or more items above the well case (Item 1), or who assigned the well case a number less than 9 on the category scale, since the instructions specifically noted that 10 is for a well day. This process eliminated 11 subjects, leaving 54 subjects who produced a total of 3,240 usable observations.

RESULTS

As Stevens and Galanter initially demonstrated [1957], the arithmetic means of category rating (on the ordinate) exhibit a concave downward relation to the geometric means of magnitude estimation (on the abscissa). Figure 1 reveals this well known concave downward relation in our data. Thus the dimension of Well-Being behaves as a prothetic continuum. The product moment correlation for this relationship is .76.

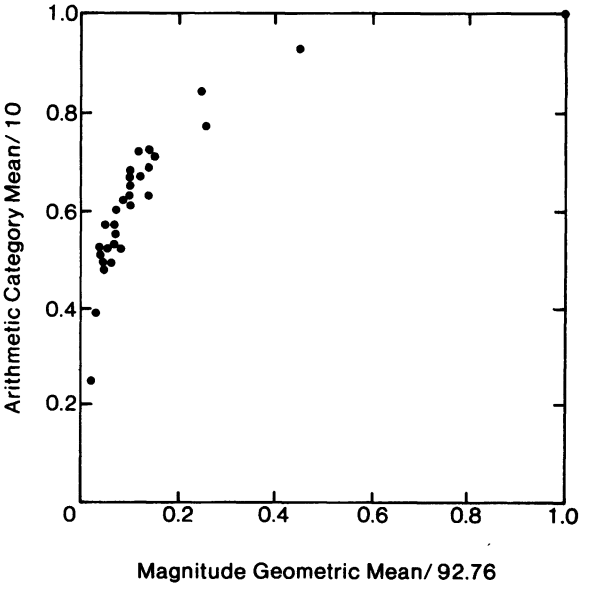

Fig. 1. Social preference ratings for 30 items representing states of dysfunction showing the classical concave downward relation between category rating and magnitude estimation.

Note that category and magnitude means have been transformed to a 0-1 scale. For category, all means were divided by 10, the top step of the scale. For magnitude, all geometric means were divided by the geometric mean weight assigned by the subjects to the Well-Day on the open-ended scale (92.76). This transforms the otherwise arbitrary numbers of the scaling procedures to a meaningful, comparable unit. The dramatic result is that all the magnitude measures of central tendency (median, arithmetic, and geometric means) compress the social preferences for almost all the items near the death state below 0.2. An item with a mean value of .72 using category rating, for example, receives a value of only .12 using magnitude estimation. If the relationship between the scaling methods is logarithmic, then a plot of category means against the logarithms of the magnitude geometric means should be approximately linear. Figure 2 demonstrates that the relationship, which has a product moment correlation of .96, is indeed approximately linear. The equation for this relation is:

$$C = .22 + .18 (\log M)$$

where

C is the arithmetic mean for the category rating for an item on a 0-1 scale, and

log M is the mean of logs (log of the geometric mean) for an item rated by magnitude estimation.

A similar comparison of the arithmetic category means versus the arithmetic magnitude means (and their logarithms) is not shown but was almost identical. This relation was apparent even when the confused and uncooperative subjects were not eliminated from the data set.

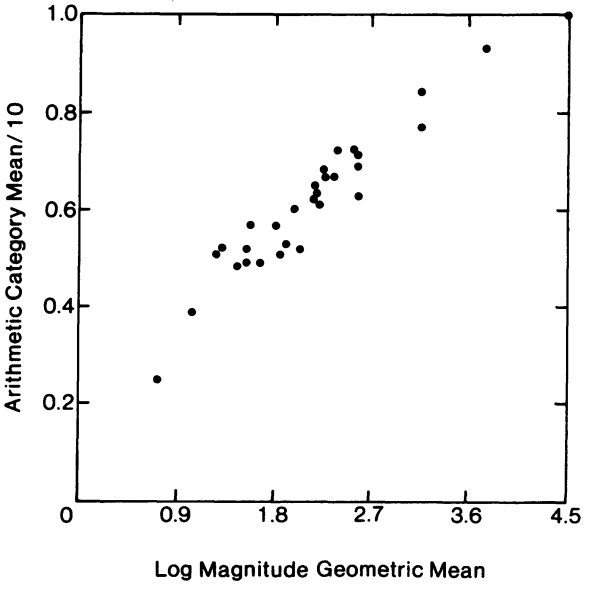

Fig. 2. Approximately linear relation of category arithmetic means to logarithms of the magnitude geometric means for items (data points) in Figure 1.

DISCUSSION

The results of this analysis confirm that social preferences or ratings of Well-Being behave as a prothetic continuum. If the continuum were metathetic, and the two methods had yielded identical results, scaling method would not be a concern for health index construction. We had originally used magnitude estimation because a fairly extensive literature held that it produced scale values with optimal properties [Stevens 1968]. The current results indicate that the scaling technique is now somewhat problematic and criteria must be established to select the best weights from those produced by different methods.

The needs of a Health Index per se are neutral in any disagreement between advocates of different scaling techniques. If magnitude estimation or a more complex technique were established as more valid, then any category data from a field survey could be transformed to yield the equivalent of the more desirable score by using a functional relation established in a careful laboratory study.

Our previous finding that the two methods agreed [Patrick et al. 1973b] was unexpected but gratifying. On logical grounds, it could be argued, either of the methods could produce an equal interval response scale. In closing the methodological loophole of the previous study, however, the non-linear relationship between the two sets of responses is now apparent. Both methods cannot be producing an equal-interval measure of preference. The results of this and subsequent research, on the other hand, do not support transforming field data from category rating to its magnitude counterpart. Figure 1 reveals that when the "ratios" from magnitude are transformed to a scale whose meaning can be interpreted directly and intuitively, the weights appear unreasonable.

Stevens was disappointed that most social scientists continued using category scales despite his repeated and vociferous objections. The major support for his magnitude estimation technique was the face validity argument that the subjects were instructed to assign their numbers "in proportion to" subjective ratios. This instruction is insufficient to establish the properties of the scale in theory [Krantz et al, 1971, p. 11], and several authors have noted Stevens' failure to provide empirical criteria for the properties that he claimed [Garner 1954; Torgerson 1960; Junge 1965; Anderson 1976].

Anderson has recently [1974, 1976] proposed a test for the equal interval property based on a simple analysis of variance. According to his functional measurement technique, the absence of a significant interaction effect in the analysis of variance establishes the equal interval property. Differences between preferences for two items which differ on only one attribute should be equal to the difference between two other items which have the same difference on that attribute. Experiments using functional measurement have demonstrated that category ratings meet this empirical criterion for the interval property while magnitude estimation does not [Anderson 1974, 1976; Weiss 1972, 1975].

Previous studies using our own case attributes have also demonstrated this absence of interaction [Patrick et al. 1973b]. One concern with the functional measurement test, which involves accepting the null hypothesis, is a possible false negative because of lack of power. In data from a probability sample of 900 San Diego households, however, this property was reconfirmed with approximately 100 subjects rating each item. Figure 3, showing data from four items, clearly demonstrates the parallelism exhibited by equal interval scales. For this analysis, both main effects were highly significant, while the F-ratio for the interaction was less than 1.0. This illustration is one from twelve similar analyses (to be reported) from balanced designs in the household survey, in which all possible interactions were non-significant.

An equal if not more important criterion for choosing between methods is whether the weights are consistent with ethical preferences [Harsanyi 1955] -- not the preferences that respondents would theoretically use for themselves, but the stated weights that they favor implementing for public policy. Our previous study [Patrick et al. 1973b] reported the only results in Health Index research (and, as far as we are aware, in social indicators research) to date using an equivalence technique which forces the trade-off among target population beneficiaries that are implied in the weighting scheme. Each of 12 comparisons among multiple groups, many composed of statewide health leaders and decisionmakers in health services, revealed non-significant differences between category rating and equivalence. The equivalence technique uses the natural social metric of the numbers of similar persons affected to provide a precisely adjustable response scale that is not biased by income, non-linearities in

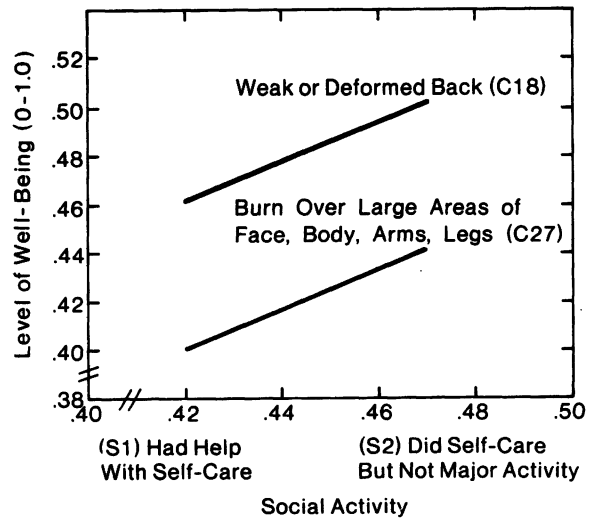

Fig. 3. Functional measurement test showing lack of interaction among items that differ by the same levels on each of two attributes (SAC and CPX) characteristic of equal interval scale.

the utility of money, prognostic or personal factors in time-tradeoffs, or aversions to gambling, which render suspect many other techniques used to measure utilities for health states. This consistency with the trade-offs implied in social choices is of major importance for the preference scale for which an equal interval measure is derived. So far as we are aware, this property has never been tested (much less demonstrated) for weights derived using magnitude estimation or any other technique in health index research.

The equal interval property may derive from the ease of administration of category scales, which means that single global ratings can be given to total case descriptions, thus considering the multiple dimensions of health states (including Symptom/Problem Complexes) jointly and simultaneously. This completely bypasses the need to rate separate attributes individually and later combine the ratings by arbitrary rules. Using such methods, the equal interval property of the total case score cannot be tested. The variance in our global ratings can be disaggregated and related to the case attributes using a simple linear model, which provides separate main-effect weights for Function Levels and Symptom/Problem Complexes and explains 96% of the variance [Chen et al. 1973].

In rejecting magnitude estimation, we do not necessarily reject all of Stevens' reservations about other attitude and preference measurement methods. In particular, we would agree that methods that unitize the dispersion in subject's responses -- just noticeable differences (jnd's) -- are not a desirable measurement unit. In our magnitude data, standard deviations increase with increasing desirability of the stimulus, but were roughly the same using the category method. The present evidence for the metric property of category responses is based, not on assumptions about stimulus or response dispersion, but on empirical tests and congruence with social choice metrics. Thus, with adequate warm-up and proper administration, category ratings apparently quantify subjective preferences directly, making later adjustments of category widths unnecessary [Blischke et al. 1975].

Torrance [1976] found that results from a time trade-off technique (of his invention) conformed to results from a version of the von Neumann-Morgenstern standard gamble better than results from category ratings. In view of its wide previous use in many circumstances, the difficulty that Torrance' subjects experienced with category rating is puzzling. This may have been because category rating was always administered at the very beginning of the interview as the first technique with very complex items.

Measures of internal reliability were not performed for category rating. In addition, the correspondence of the category rating and the time trade-off technique to the standard gamble were tested on only six items clustered near the middle of the scale. Such a study does not seem to justify Torrance's conclusion that category rating is inadequate for health index construction.

Unfortunately, even magnitude estimation does not offer the opportunity to incorporate the unbounded concept of "positive" mental health states in a health index. That limitation is due to the lack of an operational (observable or reportable) definition of the "positive" attribute to which utilities can be assigned. If it were possible to say that some persons had "positive" health attributes, while others did not, then the presence of the attribute(s) could be incorporated in the state of optimum function weighted 1.0, and the absence of the attribute(s) would simply be scored lower.

Although this would depress all values on the 0-1 scale, the scale would have been altered by incorporating a higher standard into the state of optimum function. The terms "positive" and "negative", in which much health and mental health jargon is couched, are totally arbitrary from an algebraic perspective. If a superior state of "positive" health were operationalized, it could be easily incorporated in the strategy of assigning consensus preferences to predefined states, regardless of the rating technique used. To the extent that such "positive" attributes affect current symptoms, problems and functioning, or prognoses, they are, of course, already reflected in the existing Index.

The demonstration of method differences should not lead to the conclusion that preference measures in health indexes are any more biased or unreliable than much health data that is currently published. All existing morbidity and mortality statistics have an implicit value component that is incompletely specified. In addition, all such specific statistics are upwardly biased as comprehensive health indicators because of the multiple other factors that they omit. The current life expectancy, for example, greatly overestimates the health status of a population because it includes no indication at all of the decreased quality of life.

Previous efforts to compensate for this lack has led to the publication of frankly subjective

data on scales such as "excellent/good/fair/poor" whose metric properties (despite high correlations with utilization, number of chronic conditions, etc.) have hardly been examined [USD/HEW, 1976, pp. 242-243]. Serious question can be raised, in fact, about even the ordinal properties of the scale [Kaplan et al. 1976], and yet its levels have frequently been treated as interval numbers in statistical models.

Almost any reasonable or approximate set of weights, applied to objectively verifiable states of function, would give a far more valid, reliable, and mathematically manipulable health indicator than aggregation of such crudely expressed individual opinions, for which the word "validity" has little if any meaning. As the science of function state classification and preference measurement progresses, actual values can be better approximated allowing consumer preferences to prevail over implicit, investigator assigned, or other ad hoc weighting procedures. Although arbitrarily weighted indexes can be shown to correlate highly with simplified versions of the IWB that omit variations at high levels of Well-Being -- the major source of IWB variance -- such numbers cannot be used to compute a meaningful weighted life expectancy which depends on precise 0-1 scale locations for the levels [Miles 1977]. Such an interpretation is essential to use a health index as a social indicator, as a tool for resource allocation, and even to quantify the health status impact of programs in evaluation research.

Anderson and his colleagues have demonstrated that the interval properties of the attribute ratings are preserved when the items include probabilities (prognoses) so the category ratings are consistent with the multiplicative properties required to treat them as expected values in decision models [Shanteau 1974, 1975; Anderson 1976]. These are precisely the properties required to compute the Weighted Life Expectancy and to estimate the output of a health program [Chen et al. 1975, Chen and Bush 1976].

In addition, all the preference distributions for the items rated were unimodal ("single-peaked"), which Black [1958] has demonstrated provides a sufficient condition to insure the transitivity of the resulting social preference function.

With the addition of the present results, our psychometric studies may be summarized as follows:

1. Preferences can be measured reliably ($r = 0.91$) from cross-validation studies using randomly created parallel forms of the procedure;
2. The values on the 0-1 scale possess equal-interval properties;
3. The category ratings are stable across different orders of testing and modes of test administration;
4. Linear statistical models accurately represent and predict ($R^2 \geq 0.96$) the mean and median global consumer ratings for individual case descriptions;
5. Age groups representing different phases of the life cycle in the case descriptions account for only about 1 percent of the variance in the preference ratings;
6. The preferences are generalizable across different social groups and their leaders, all of whom seem to share a consensus on the terminal values associated with the Function Levels; and
7. The category ratings are consistent with results from procedures designed to test for the ethical preferences implied in social choices, and have unimodal distributions which insure social transitivity.

With data now available, we will soon be able to examine the stability of the mean and median preferences over time.

This accumulation of evidence supports the notion that category ratings give social preference weights that are as nearly valid and with as desirable properties as any other techniques tried to date. Contrary to previous suggestions [Arrow 1963, Stevens 1966], magnitude estimation does not appear appropriate as a measurement method for a health status index and is probably inappropriate also for social indicators [Sellin and Wolfgang 1964] and other criteria of social choice.

REFERENCES

Anderson NH, Algebraic Models in Perception. In EC Carterette and MP Friedman, eds., HANDBOOK OF PERCEPTION, V. 2. NY: Academic Press, 1974, 215-291.

Anderson NH, How Functional Measurement Can Yield Validated Interval Scales of Mental Quantities. J APPL PSYCHOL 61:677-692, 1976.

Arrow KJ, SOCIAL CHOICE AND INDIVIDUAL VALUES. New Haven: Yale Univ. Press, 1963.

Black D, THE THEORY OF COMMITTEES AND ELECTIONS. Cambridge: University Press, 1958.

Blischke WR, Bush JW and Kaplan RM, A Successive Intervals Analysis of Social Preference Measures for a Health Status Index. HEALTH SERV RES 10(2):181-198, 1975.

Bush JW, Chen Milton and Patrick DL, Cost-Effectiveness Using a Health Status Index: Analysis of the New York State PKU Screening Program. In R Berg, ed., HEALTH STATUS INDEXES. Chicago: Hospital Research and Educational Trust, 1973, 172-208.

Chen Milton and Bush JW, Maximizing Health System Output with Political and Administrative Constraints Using Mathematical Programming. INQUIRY 13(3):215-227, Sept 1976.

Chen Milton, Bush JW and Patrick DL, Social Indicators for Health Planning and Policy Analysis. POLICY SCIENCES 6(1):71-89, 1975.

Chen Milton, Bush JW, Patrick DL and Blischke WR, Statistical Models of Social Preferences for Constructing a Health Status Index. Springfield, VA: National Technical Information Service, Pub. No. PB 236 155/8ST, 1973.

Garner WR, Context Effects and the Validity of Loudness Scales. J EXP PSYCHOL 48:218-224, 1954.

Harsanyi JC, Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility. J POLIT ECON 63(3):309-321, 1955.

Junge K, SOME PROBLEMS OF MEASUREMENT IN PSYCHOPHYSICS: A THEORETICAL STUDY. Oslo, Norway: Scandinavian University Books, 1966.

Kaplan RM, Bush JW and Berry CC, Health Status: Types of Validity for an Index of Well-Being. HEALTH SERV RES 11(4):478-507, 1976.
Krantz DH, Luce RD, Suppes P and Tversky A, FOUNDATIONS OF MEASUREMENT. NY: Academic Press, 1971.
Miles DL, Health Care Evaluation Project Terminal Progress report, National Center for Health Services Research Grant 5 R01 HS 01568, July 1977.
Patrick DL, Bush JW and Chen Milton, Toward an Operational Definition of Health. J HEALTH SOC BEHAV 14(1):6-23, 1973a.
Patrick DL, Bush JW and Chen Milton, Methods for Measuring Levels of Well-Being for a Health Status Index. HEALTH SERV RES 8(3):228-245, 1973b.
Sellin T and Wolfgang ME, THE MEASUREMENT OF DELINQUENCY. NY: Wiley, 1964.
Shanteau JC, Component Processes in Risky Decision Making. J EXP PSYCHOL 103:680-691, 1974.
Shanteau JC, An Information Integration Analysis of Risky Decision Making. IN MF Kaplan and S Schwartz, eds., HUMAN JUDGMENT AND DECISION PROCESSES. NY: Academic Press, 1975.
Stevens SS, Issues in Psychophysical Measurement. PSYCHOL REV 78:426-450, 1971.
Stevens SS, A Metric for the Social Consensus. SCIENCE 151:530-541, Feb 1966.
Stevens SS, Perceptual Magnitude and Its Measurement. In EC Carterette and MP Friedman, eds., HANDBOOK OF PERCEPTION, V. 2. NY: Academic Press, 1974, 361-387.
Stevens SS, Ratio Scales of Opinion. In DK Whitla, ed., HANDBOOK OF MEASUREMENT AND ASSESSMENT IN BEHAVIORAL SCIENCES. Reading, Mass.: Addison-Wesley, 1968, 171-199.
Stevens SS and Galanter E, Ratio Scales and Category Scales for a Dozen Perceptual Continua. J EXP PSYCHOL 54:377-411, 1957.
Torgerson WS, Quantitative Judgment Scales. In Gulliksen and Messick, eds., PSYCHOLOGICAL SCALING: THEORY AND APPLICATIONS. NY: Wiley, 1960.
Torrance GW, Social Preferences for Health States: An Empirical Evaluation of Three Measurement Techniques. SOCIO-ECON PLANN SCI 10:129-136, 1976.
USDHEW, Public Health Service, HEALTH, UNITED STATES, 1975. Rockville, MD: USDHEW Pub. No. (HRA) 76-1232, 1976.
Weiss DJ, Averaging: An Empirical Validity Criterion for Magnitude Estimation. PERCEPTION AND PSYCHOPHYSICS 12:385-388, 1972.
Weiss DJ, Quantifying Private Events: A Functional Measurement Analysis of Equisection. PERCEPTION AND PSYCHOPHYSICS 17:351-357, 1975.

APPENDIX I: SCALES AND DEFINITIONS FOR CLASSIFICATION OF FUNCTION LEVELS*

MOBILITY

5 Drove car and used bus or train without help
4 Did not drive, or had help to use bus or train
3 In house
2 In hospital
1 In special care unit

PHYSICAL ACTIVITY

4 Walked without physical problems
3 Walked with physical limitations
2 Moved own wheelchair without help
1 In bed or chair

SOCIAL ACTIVITY

5 Did work, school or housework, and other activities
4 Did work, school, or housework, but other activities limited
3 Limited in amount or kind of work, school, or housework
2 Performed self-care, but not work, school, or housework
1 Had help with self-care activities

* Instruments for classification of persons into one and only one Function Level for multiple days available from the authors