MEDICAL CARE
November 1989, Vol. 27, No. 11

Communication

Interday Reliability of Function Assessment for a Health Status Measure

The Quality of Well-Being Scale

JOHN P. ANDERSON, PHD,* ROBERT M. KAPLAN, PHD,* CHARLES C. BERRY, PHD,* JAMES W. BUSH, MD,* AND RUBEN G. RUMBAUT, PHD†

Interview-based measures of functional health status[1-7] may play a significant role in the evaluation of health policies and treatments. As one of a series of methodologic reports, this paper focuses on the interday reliability of information obtained by the Quality of Well-Being (QWB) scale.

The Quality of Well-Being Scale

The QWB scale combines preference-weighted measures of symptoms and functioning to provide a numeric point-in-time expression of well-being, which ranges from zero (0) for death to one (1.0) for asymptomatic optimum functioning. Table 1 presents 23 Symptom/Problem Complexes (CPX) along with their preference weights. Utilization of this CPX list does not require any assumptions about the intensity or duration of symptoms and problems, or about the underlying pathology, if any. The measure simply indicates the symptom's presence or absence on a given day.

The QWB 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).

The development of the QWB has been described in several publications.[3-8] Validation information about the QWB scale has been reported earlier,[8] as has information on the accuracy of function classification methods,[9] the influence of various measure-

Support provided for the following studies by these sources: the follow-up survey by grant no. 2 R18 HS00702 from the National Center for Health Services Research, the burn study by contract no. 240-77-0158 from DHHS, the Indochinese Health and Adaptation Research Project by Grant No. 1 R01 HD15699 from the National Center for Child Health and Human Development, the COPD Project by grant nos. K04 HL 00809 and R01 HL 34732, and the Diabetes Project by grant no. R01 AM 27901.

* From the Department of Community and Family Medicine, University of California at San Diego, La Jolla, California.

† From the Department of Sociology, San Diego State University, San Diego, California.

Address correspondence to: John P. Anderson, PHD, Department of Community and Family Medicine, University of California at San Diego, La Jolla, CA 92093.

TABLE 1. List of Quality of Well-Being Scale Symptom/Problem Complexes (CPX) With Calculating Weights

CPX No.	CPX Description	Weights
1	Death (not on respondent's card)	−0.727
2	Loss of consciousness such as seizure (fits), fainting, or coma (out cold or knocked out)	−0.407
3	Burn over larger areas of face, body, arms, or legs	−0.367
4	Pain, bleeding, itching, or discharge (drainage) from sexual organs; does not include normal menstrual (monthly) bleeding	−0.349
5	Trouble learning, remembering, or thinking clearly	−0.340
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	−0.333
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	−0.299
8	Pain, burning, bleeding, itching, or other difficulty with rectum, bowel movements, or urination (passing water)	−0.292
9	Sick or upset stomach, vomiting or loose bowel movements, with or without fever, chills, or aching all over	−0.290
10	General tiredness, weakness, or weight loss	−0.259
11	Cough, wheezing, or shortness of breath with or without fever, chills, or aching all over	−0.257
12	Spells of feeling upset, being depressed, or of crying	−0.257
13	Headache, or dizziness, or ringing in ears, or spells of feeling hot, nervous, or shaky	−0.244
14	Burning or itching rash on large areas of face, body, arms, or legs	−0.240
15	Trouble talking, such as lisp, stuttering, hoarseness, or inability to speak	−0.237
16	Pain or discomfort in one or both eyes (such as burning or itching), or any trouble seeing after correction	−0.230
17	Overweight for age and height or skin defect of face, body, arms, or legs, such as scars, pimples, warts, bruises, or changes in color	−0.186
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)	−0.170
19	Taking medication or staying on a prescribed diet for health reasons	−0.144
20	Wore eyeglasses or contact lenses	−0.101
21	Breathing smog or unpleasant air	−0.101
22	No symptoms or problem (not on respondent's card)	−0.000
23	Standard symptom/problem (not on respondent's card)	−0.257

ment problems on the size, direction, and effects of function misclassification,[10] and the stability and generalizability of the utility weights used for the QWB.[11] In this report we focus on the interday reliability of the measure.

Methods

Empirical Studies

This paper uses data from five empirical studies. Each will be identified in terms of 1) subject population, 2) number of QWB days involved, 3) number of respondents/subjects, and 4) languages involved, if other than English.

Follow-up Survey. This was a one-year follow-up of a probability sample of respondents, selected children, and dysfunctional persons in San Diego county. The sampling characteristics have been extensively described elsewhere.[8,9] Briefly, a probability sample of 1,025 survey subjects from metropolitan San Diego were involved, with a total of 8 QWB days for each respondent or subject. The study included oversampling of dysfunctional persons, with appropriate representations of various ages, ethnic groups, and disabilities.

Burn Study. This was a clinical follow-up of adult patients who had been treated at the Regional Burn Treatment Center of the UCSD Medical Center, contacted when

TABLE 2. Quality of Well-Being Scale Elements and Calculating Formulas

Step No.	Step Definition	Weight
Mobility Scale (MOB)		
5	No limitations for health reasons	−0.000
4	Did not drive a car, health related (younger than 16); did not ride in a car as usual for age, 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	−0.062
2	In hospital, health related	−0.090
Physical Activity Scale (PAC)		
4	No limitations for health reasons	−0.000
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	−0.060
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	−0.077
Social Activity Scale (SAC)		
5	No limitations for health reasons	−0.000
4	Limited in other role activity, health related	−0.061
3	Limited in major (primary) role activity, health related	−0.061
2	Performed no major role activity, health related, but did perform self-care activities	−0.061
1	Performed no major role activity, health related, and did not perform or had more help than usual in performance of one or more self-care activities, health related	−0.106

Calculated Formulas

Formula 1: Point-in-time Well-Being score for an individual (W):

$$W = 1 + (CPXwt) + (MOBwt) + (PACwt) + (SACwt)$$

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 follows:

QWB Element	Description	Weight
CPX-11	Cough, wheezing, or shortness of breath, with or without fever, chill, or aching all over	−0.257
MOB-5	No limitations	−0.000
PAC-1	In bed, chair, or couch for most or all of the day, health related	−0.077
SAC-2	Performed no major role activity, health related, but did perform self-care activities	−0.061

$$W = 1 + (-0.257) + (-0.000) + (-0.077) + (-0.061) = 0.605$$

Formula 2: General Health Policy Model Formula for Well-Years (WY) as an output measure:

$$WY = [\text{No. of Persons} \times (CPXwt + MOBwt + PACwt + SACwt)] \times \text{Time}$$

they were at or past the point of maximum recovery from burn injury. A total of 145 patients were involved, with 4 QWB days for each patient.

Indochinese Health and Adaptation Research Project. This was a survey, with a one-year follow-up, of a probability sample of Indochinese refugee respondents living in San Diego county. QWB interviews were done mainly in translation to the refugee's native language, although a few were completed in English. The languages involved were Chinese (three dialects: Mandarin, Cantonese, and Chao-Chao), Hmong (a minority people from Laos), Khmer (Cambodian), and Vietnamese. A total of 599 respondents were involved in the first survey, with 500 of them reinterviewed at follow-up. In each interview, 6 QWB days were involved.

Chronic Obstructive Pulmonary Disease Project. This study compared rehabilitation with education interventions for older adults with Chronic Obstructive Pulmonary Disease (COPD). These patients experience significant physical dysfunction as a result of moderate to severe respiratory abnormalities. Six-day QWB scores are to be taken from each of 120 patients on five separate occasions over two years. As the study is still in progress, only preliminary data on less than the full sample and full number of interviews can be reported.

Diabetes Project. The 76 participants in this small clinical trial were afflicted with NonInsulin Dependent Diabetes Mellitus (NIDDM). This group, typically characterized by obesity, may benefit significantly from weight loss. The trial compared the efficacy of diet, exercise, diet and exercise, or education control with weight loss and QWB scores. QWB information was obtained about 4 days at the point before treatment, after 3 months, 6 months, 12 months, and 18 months.

Reliability Calculations. Two methods are usually used to estimate interday reliability of information on function developed by health measures. For numeric scores, Pearson product-moment correlation coefficients are calculated for each successive dyad of days, and for specific reports of dysfunction, an Agreement Percent (AP) is calculated to reflect day-by-day agreement in reports of dysfunction, where a report of dysfunction on both days of a pair represent agreement, and a report of dysfunction on one of the days (the other reporting full function) represents disagreement. The formula for calculating the AP is:

$$\text{AP} = \text{Number of Agreements}/(\text{Number of Agreements} + \text{Number of Disagreements})$$

An AP for each group represents the average of APs for each subject. As the QWB employs no items, no attempt will be made to calculate item APs.

Where interday reliability of a measure is being explored, the usual situation is one of test-retest; that is, having the same measure administered at two different points in time. In the case of the QWB, function information about multiple contiguous days in the immediate past is elicited and recorded.

Given that multiple days are involved, the number of dyads reported will be number of days minus one. Thus, with a four-day QWB, there are three dyads, Day 4-Day 3, Day 3-Day 2, and Day 2-Day 1. In the case of the follow-up survey, for example, we have 7 dyads × 1,025 subjects, or 7,175 comparisons. All told, in the Agreement Percent analysis, which includes the follow-up survey, the burn study, and the Indochinese surveys, 13,100 person-dyad comparisons will be involved.

The correlation analysis involves all of the people itemized above, and in addition will include 192 interviews with COPD patients (192 × 5 dyads = 960 comparisons) and 70 interviews with Diabetes patients (70 × 3 dyads = 210), for a total of 14,270 comparisons.

Results

Table 3 reports the interday correlation coefficients for all studies. These correlations ranged from a low of 0.78 to a high above 0.99, with most of the coefficients being above 0.90, and only one below 0.80. Table 4 reports the Agreement Percent figures for the available studies. These run from a low of 0.77 to a high of 1.0, with most being above 0.80.

Discussion

The results shown in Table 3 are relatively constant across the available linguistic variations—QWB reliability appearing as good in Chinese, Hmong, Khmer, and Vietnamese as it is in English. The reliabilities are somewhat higher (0.9 or above) in populations we know to be impaired (burn patients, COPD and diabetes patients), so the instrument appears to be measuring dysfunction with a high degree of descriptive reliability.

These interday correlations, while high, are in our view not the most important evidence for the reliability of the QWB. The correlations involve scores for symptoms and problems also, which are not part of the analysis. Indeed, as we have previously demonstrated,[9] where standards for scientific instruments are concerned, correlations can be grossly insensitive as measures of instrument quality.

A multiple-day (meaning four or more) snapshot of health status provides a more useful window for reliability analysis than can be provided by a two-day period. In this paper, we examine how reasonable it might be to consider variation within the multiple-day window to be measurement error in the characterization of current well-being.

The data from these analyses tend to support the reliability of the QWB. If the AP of 0.8 or 0.9 indicates the amount of measurement error present, it means the reliability of this health measure is substantial. If an AP of 0.8 or 0.9 indicates the true empiric

TABLE 3. Interday Quality of Well-Being Score Correlations by Study, Day, and Sample Status

	Day 8-Day 7	Day 7-Day 6	Day 6-Day 5	Day 5-Day 4	Day 4-Day 3	Day 3-Day 2	Day 2-Day 1
1975 Follow-Up survey							
Respondents (No. = 681)	0.96	0.95	0.93	0.91	0.91	0.91	0.92
Selected children (No. = 274)	0.93	0.78	0.87	0.87	0.87	0.91	0.92
Dysfunctionals (No. = 70)	0.99	0.99	0.98	0.94	0.97	0.96	0.97
Indochinese project							
T1 Respondents (No. = 598)			0.94	0.92	0.90	0.89	0.90
T2 Respondents (No. = 500)			0.95	0.94	0.94	0.92	0.93
Burn study							
Respondents (No. = 143)					0.83	0.87	0.83
COPD project respondents							
First interview (No. = 84)			0.98	0.84	0.88	0.92	0.95
Second interview (No. = 63)			0.90	0.81	0.95	0.88	0.92
Third interview (No. = 45)			0.98	0.95	0.93	0.90	0.80
Diabetes project							
First interview (No. = 70)					0.96	0.97	0.94

TABLE 4. Interday Quality of Well-Being Agreement Percent Reliability by Study, Day, and Sample Status

	Day 8-Day 7	Day 7-Day 6	Day 6-Day 5	Day 5-Day 4	Day 4-Day 3	Day 3-Day 2	Day 2-Day 1
1975 Follow-Up survey							
Respondents (No. = 681)	0.82	0.85	0.82	0.78	0.82	0.86	0.85
Selected children (No. = 274)	0.87	0.79	0.79	0.80	0.77	0.88	0.86
Dysfunctionals (No. = 70)	1.00	1.00	1.00	0.87	0.95	0.88	0.86
Indochinese project, T-1							
Respondents (No. = 598)			0.94	0.84	0.84	0.89	0.88
Indochinese project, T-2							
Respondents (No. = 500)			0.92	0.81	0.80	0.90	0.90
Burn study							
Respondents (No. = 145)					0.97	0.97	0.98

situation of changes between function and dysfunction over time that people normally experience, and since the AP for the QWB comes in this range, this also could be evidence of the instrument's accuracy and reliability.

We suspect that the discrepancy between these AP figures and 1.0 is not the unreliability of the instrument, but the actual changes in dysfunction that people experience over time. However, no "gold standard" means of proving this is available.

AP is, as was pointed out by Pollard et al.,[12] a conservative measure of instrument reliability, and we hold that this is the most important evidence for reliability of the QWB instrument. Some evidence for the latter hypothesis comes from the COPD and diabetes trials. In these cases, we would expect day-to-day variation to be small. However, the sensitivity of the instrument is reflected in the ability to detect relatively small improvements in function. In the COPD study, the QWB detected small changes over the course of time in dysfunction attributable to behavioral interventions, and these changes corresponded to other measures such as exercise tolerance and compliance. In the diabetes study, changes in QWB were correlated with changes on other measures of diabetes control.

In all of our studies, it has been observed that reports of dysfunction tend to occur in blocks of days, where persons with acute diseases start off functional, become dysfunctional for a few days, then return to full function. By contrast, persons with more chronic diseases tend to start off dysfunctional and stay dysfunctional. When persons report multiple dysfunctions, the dysfunctions tend to track together: if one dysfunction happens on days 3, 4, and 5, other dysfunctions are likely to occur among those days, and not in complete isolation from one another.

One possible problem with the reported analyses is that reliability estimates may have been inflated because of a "methods effect." Reliability from one day to the next was ascertained in a retrospective manner for several days simultaneously. Ideally, data would have been obtained independently on each day (See Appendix A). However, there is also substantial evidence that acute variations in health are adequately tapped with the retrospective method. For example, responents who experienced acute illnesses showed decrements in QWB for the specific days when they were most sick. There is day-to-day variation within each of the patient groups. Although reliability may be somewhat inflated by the retrospective assessment for several days, there is little evidence that respondents simply report the same information for each day.

A recent clinical trial evaluating oral gold (Auranofin) for patients with rheumatoid arthritis suggested that the QWB was among the most sensitive indicators of clinical change. The point at which the placebo and treatment groups began to diverge in terms of QWB scores corresponded to the point at which the medication was expected to reach clinically effective levels in blood.[30] For chronic illness, QWB reports of dysfunction appear to be relatively stable over closely spaced points in time, but are also sensitive to changes in the conditions.

Finally, it is important to emphasize that the preference weights are assigned by a linear model. Tables 1 and 2 are constructed so that the weights for symptoms were entered first in a step-wise model of preference. This must be taken into consideration when interpreting results. For example, a weight of −0.101 is given for breathing smog or unpleasant air. This is about the same weight given for someone who performs no major role activity in the Social Activity Scale. However, it is not possible to have the functional limitation (Table 2) without having at least one symptom. Thus, a person who performs no role activity because they breathed unpleasant air would be given a score of (1.0 − 0.101 − 0.106 = 0.793). It is unlikely that the person would be limited in their role activity because of breathing unpleasant air. It is much more likely that this limitation would be associated with a more severe symptom, such as missing limbs. In addition, limitations in function are usually associated with limitations on the other scales. Thus, a typical score for someone who performed no major activity would be about 0.48. The score for someone who has no functional limitations but breathes bad air would be 0.90.

In summary, we present new evidence for the interday reliability of the QWB from several different populations. We find substantial evidence to support the use of this measure in clinical research.

References

1. Bergner M, Bobbitt RA, with Carter WB, et al. The sickness impact profile: development and final revision of a health status measure. Med Care 1981;19:787.

2. Stewart AL, Ware JE Jr., Brook RH. Advances in the measurement of functional status: construction of aggregate indexes. Med Care 1981;19:473.

3. Bush JW, Fanshel S. Measuring health system output using a health status index. In: Hopkins C, ed. Outcomes conference I–II. USD/HEW, Public Health Service, Health Services and Mental Health Administration. Washington, DC: 1969.

4. Bush JW, Fanshel S, Chen M. Analysis of a tuberculin testing program using a health status index. Journal of Socio-Economic Planning Sciences 1972;6:49.

5. Bush JW, Chen M, Patrick DL. Cost-effectiveness using a health status index: analysis of the New York State PKU screening program. In: Berg R, ed. Health Status Indexes. Chicago: Hospital Research and Educational Trust, 1973:172.

6. Bush JW. General health policy model/Quality of Well-Being (QWB) scale. In: Wenger NK, et al, eds. Assessment of quality of life in clinical trials of cardiovascular therapies. New York: Le Jacq, 1984.

7. Fanshel S, Bush JW. A health status index and its application to health services outcomes. Operations Research 1970;18:1021.

8. Kaplan RM, Bush JW, Berry CC. Health status: types of validity and the index of well-being. Health Services Research 1976;11:478.

9. Anderson JP, Bush JW, Berry CC. Classifying function for health outcome and Quality-of-Life evaluation: self- versus interviewer modes. Med Care 1986;24:454.

10. Anderson JP, Bush JW, Berry CC. Internal consistency analysis: a method for studying the accuracy of

function assessment for health outcome and quality of life evaluation. J Clin Epidemiol 1988;41:127.

11. Kaplan RM, Bush JW, Berry CC. The reliability, stability and generalizability of a health status index. In: Proceedings of the American Statistical Association, Social Statistics Section, 1978:704.

12. Pollard WE, Bobbitt RA, Bergner M, et al. The Sickness Impact Profile: reliability of a health status measure. Med Care 1976;14:146.

13. Anderson JP, Bush JW, Chen M, Dolenc D. Policy space areas and properties of Benefit-Cost/Utility analysis. JAMA 1986;255:794.

14. Bush JW, Fanshel S, Chen M. Analysis of a tuberculin testing program using a health status index. Journal of Socio-Economic Planning Sciences 1972;6:49.

15. Bush JW, Chen M, Patrick DL. Cost-effectiveness using a health status index: analysis of the New York State PKU screening program. In: Berg R, ed. Health Status Indexes. Chicago: Hospital Research and Educational Trust, 1973:172.

16. Epstein KA, Schneiderman LJ, Bush JW, Zettner A. The 'abnormal' screening serum thyroxine (T4): analysis of physician response, outcome, cost and health effectiveness. Journal of Chronic Diseases 1981;34:175.

17. Anderson JP, Moser RJ. Parasite screening and treatment among Indochinese refugees: cost-benefit/utility and the General Health Policy Model. JAMA 1985;253:2229.

18. Chen M, Bush JW. Maximizing health system output with political and administrative constraints using mathematical programming. Inquiry 1976; 13:215.

19. Chen Milton, Bush JW, Patrick DL. Social indicators for health planning and policy analysis. Policy Sciences 1975;6:71.

20. Bush JW, Blischke WR, Berry CC. Health indices, outcomes, and quality of care. In: Yaffe R, Zalkind D, eds. Evaluation of Health Services Delivery. New York: Engineering Foundation, 1975:313.

21. Patrick DL, Bush JW, Chen M. Method for measuring levels of well-being for a health status index. Health Services Research 1973;8:228.

22. Patrick DL, Bush JW, Chen M. Toward an operational definition of health. Journal of Health and Social Behavior 1973;14:6.

23. Kaplan RM, Bush JW, Berry CC. Health status index: category rating versus magnitude estimation for measuring levels of well-being. Med Care 1979;17:501.

24. Bush JW, Anderson JP, Kaplan RM, Blischke WR. 'Counterintuitive' preferences in health related quality of life measurement. Med Care 1982;20:516.

25. Bush JW, Chen M, Zaremba J. Estimating health program outcomes using a Markov equilibrium analysis of disease development. Am J Public Health 1971;61:2362.

26. Berry CC, Bush JW. Estimating prognoses for a dynamic health index, the weighted life expectancy, using the multiple logistic with survey and mortality data. In: Proceedings of the American Statistical Association, Social Statistics Section, 1978:716.

27. Anderson JP, Bush JW, Berry CC. Performance versus capacity: a conflict in classifying function for health status measurement. Washington, DC, 1977. (Presented at APHA annual meetings.)

28. Kaplan RM, Atkins CJ, Timms RM. Validity of a Quality of Well Being scale as an outcome measure in chronic obstructive pulmonary disease. Journal of Chronic Diseases 1984;37:85.

29. Kaplan RM, Bush JW. Health related quality of life measurement for evaluation research and policy analysis. Health Psychology 1982;1:61.

30. Bombardier C, Ware J, Russell IJ, et al. Auranofin therapy and quality of life in patients with rheumatoid arthritis: results of a multicenter trial. Am J Med 1986;81:565.

Appendix A. Reliability, Stochastic Change, and Error

Symbolically, a measure of health status (W) might be represented as

$$W = x + \delta + E$$

where W is a measure of wellness, δ is a transient aspect of wellness, and E is error.

The variability of W can be decomposed as follows:

$$\sigma_W{}^2 = \sigma_x{}^2 + \sigma_\delta{}^2 + \sigma_E{}^2$$

The reliability of W is

$$\frac{\sigma_x{}^2 + \sigma_\delta{}^2}{\sigma_x{}^2 + \sigma_\delta{}^2 + \sigma_E{}^2}$$

In most situations, the transient component (δ) is unknown or unmeasurable. Thus, the reliability estimate will be attenuated by a factor ρ_{12}. The attenuated reliability, r*, is

$$r^* = \frac{\sigma_x{}^2 + \rho_{12}\sigma_\delta{}^2}{\sigma_x{}^2 + \sigma_\delta{}^2 + \sigma_E{}^2}$$

Whenever $\rho_{12} < 1$, $r^* < r$.

In other words, failure to consider the transient component will underestimate true reliability.