The Quality of Well-Being Scale: critical similarities and differences with SF-36

ROBERT M. KAPLAN, THEODORE G. GANIATS, WILLIAM J. SIEBER AND JOHN P. ANDERSON

Department of Family and Preventive Medicine, University of California, CA, USA

Abstract

Purpose. To summarize the development and application of a generic measure of health-related quality of life known as the Quality of Well-Being Scale (QWB).

Background. The QWB is part of a general health policy model. The measure includes functional components for mobility, physical activity, and social activity. In addition, it includes a comprehensive list of symptoms and problems. QWB scoring allows placement of each individual on a continuum of wellness ranging from 0 (for dead) to 1.0 for asymptomatic full function. The General Health Policy Model combines this point in time measure with information on diagnosis and mortality to estimate quality-adjusted life-years (QALYs). Forming a ratio of program costs to QALYs yields estimates of cost per quality-adjusted life year. Evidence supports the validity of the QWB for a wide variety of applications in population monitoring, descriptive studies of patient populations, and clinical trials. We offer a variety of comparisons between the QWB and Medical Outcomes Study Short Form 36 (SF-36).

Main findings. In comparison with the SF-36, the QWB provides less information on health profiles, but has the advantage of providing a metric that can be used for cost-utility or cost-effectiveness analysis. Although ceiling effects are common for some SF-36 subscales, perfect scores on the QWB are very rare. The QWB has an approximately normal distribution for populations of adults. Although, often criticized for not including a mental health component, we present evidence documenting the validity of the QWB for patients with psychological and psychiatric diagnoses.

Conclusions. The QWB and SF-36 are alternative comprehensive measures of health outcomes.

Keywords: outcomes assessment, outcomes research, quality of life, Quality of Well-Being Scale (QWB), SF-36

A variety of different methods are now available to measure health-related quality of life. Many of these measures are specific to particular illnesses or diseases. However, there is a continuing need for general or generic measures that can be used for population monitoring, evaluation research, individual clinical decisions, or as outcome measures in randomized clinical trials. The purposes of this paper are to: (i) review the Quality of Well-Being Scale (QWB) and (ii) compare the QWB with the Medical Outcomes Study 36 Item Short Form (SF-36).

QWB—general health policy model

General background

The General Health Policy Model grew out of substantive theories in economics, psychology, medicine, and public health. The model of health status includes components for mortality (death) morbidity (health-related quality of life) and time. The rationale for the model is that diseases and disabilities are important for two reasons. First, illness may cause the life expectancy to be shortened. Second, illness may make life less desirable at times prior to death (health-related quality of life) [1–5].

Central to the General Health Policy Model is a general conceptualization of quality of life. The QWB is a method of measuring quality of life for calculations in the model. The QWB is a preference-weighted measure combining three scales of functioning with a measure of symptoms and problems to produce a point-in-time expression of well-being that runs from 0 (for death) to 1.0 (for asymptomatic full function) [3]. The model separates aspects of health status and life quality into distinct components. These are life expectancy (mortality), functioning and symptoms (morbidity), preference for observed functional states (utility) and...
duration of stay in health states (prognosis). In addition to classification into observable levels of function, individuals are also classified by symptoms or problems. Symptoms, such as fatigue or a sore throat might not be directly observable by others, while problems, such as a missing limb might be noticeable by others. On any particular day, nearly 80% of the general population is optimally functional. However, over an interval of 8 consecutive days, only 12% experience no symptoms [3]. Symptoms or problems may be severe, such as serious joint pain, or minor such as taking medication or following a prescribed diet for health reasons.

In order to obtain preference weights for observable health states, peer judges place the observable states of health and functioning onto a preference continuum ranging from 0 for death to 1.0 for asymptomatic full function [4-6]. In addition to the morbidity component, the model requires mortality data as from life tables [7], direct measurement [8], or clinical experience. The quality-adjusted life expectancy is the current life expectancy adjusted for diminished quality of life associated with dysfunctional states and the durations of stay in each state. The model quantifies the health activity or treatment program in terms of the quality-adjusted life-years that it produces or saves.

Reliability

Reliability is an estimate of the proportion of variance in a test or measure that is true score. Subtracting this proportion from 1.0 gives an estimate of the proportion of variation assumed to be error. There are several potential sources of error. Two important sources are item sampling and time sampling. One of the basic tenets of psychometric theory is that each item in a test or measure is an unbiased and representative sample from the domain under study. In the construction of an intelligence test, for example, there is an infinite number of items that might represent intellectual ability. The Domain Sampling Model [9] assumes that items are sampled from this domain of performance and that each item is assumed to be an unbiased estimate of the underlying trait. Reliability is estimated from the inter-item correlations and the extent to which the items are intercorrelated characterizes the reliability of the measure. Reliable measures are those for which the component items measure the same construct.

Measures of health status may be derived from a very different theoretical model. Items may not be considered random samples from a large domain because each may have a very specific meaning. For example, report of a severe headache offers very specific information. The item is not randomly sampled from all possible symptoms and the meaning of reporting a headache is very different from the meaning of reporting difficulty urinating. Thus, item sampling, as known in psychometric theory, is not necessarily relevant to many health status measures. A question about a headache, as used in clinical medicine, is not randomly selected from a large domain of questions about health.

A second type of reliability in psychometric theory concerns time sampling. Psychological traits are considered to be stable over the course of time. If Sally is intelligent today, we expect her to be equally intelligent 2 weeks from today. Variation in her performance across the 2 weeks might be attributable to measurement error. Thus, test–retest estimates are an important source of information.

This same logic may not apply to health status measures. If Sally is very sick today, we may not expect her to be equally sick in 2 weeks time. Differences between health scores taken at two points in time may mean that she recovered from her illness or that she got sicker. When the underlying construct is expected to change over time, test–retest evaluations may have very little meaning. As a result, traditional reliability data have less meaning for the QWB.

We are not implying that utility measures are exempt from reliability assessment. Measurement error is assessed in several other ways. For example, we consider reliability assessed over short intervals when large swings in health status are not expected. In addition, we can consider agreement on classification of functioning using different methods.

Considerable evidence shows that the QWB scores assessed on consecutive days are highly reliable. Table 1 summarizes consecutive day QWB correlations for a variety of populations. As the table shows, the measure has good short term stability [10]. A second method for calculating interday reliability estimated agreement per cent for reports of dysfunction on consecutive days. The agreement percentage (AP) is calculated as:

\[ AP = \frac{\text{no. of agreements}}{\text{no. of agreements} + \text{no. of disagreements}} \]

These values are shown in the righthand column of Table 1. The values were not available for three of the seven populations. The reported values are from Anderson et al. [10].

<table>
<thead>
<tr>
<th>Population</th>
<th>Reliability</th>
<th>Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>General population adults ((n = 681))</td>
<td>0.96</td>
<td>0.82</td>
</tr>
<tr>
<td>General population children ((n = 274))</td>
<td>0.93</td>
<td>0.87</td>
</tr>
<tr>
<td>Adults in California—Indochinese community ((n = 598))</td>
<td>0.94</td>
<td>0.94</td>
</tr>
<tr>
<td>Burn patients ((n = 143))</td>
<td>0.83</td>
<td>0.97</td>
</tr>
<tr>
<td>Non-head trauma patients(^2) ((n = 1048))</td>
<td>0.90</td>
<td></td>
</tr>
<tr>
<td>Chronic obstructive pulmonary disease patients(^2) ((n = 84))</td>
<td>0.98</td>
<td></td>
</tr>
<tr>
<td>Diabetes patients(^2) ((n = 70))</td>
<td>0.96</td>
<td></td>
</tr>
</tbody>
</table>

\(^1\)Data from [10] and JP Anderson, unpublished work. 
\(^2\)Agreement data not available for trauma, COPD and diabetes patients.
which, in turn, had less impact than emphysema [7]. As chronic obstructive pulmonary disease [13], AIDS disease was shown to be a less serious problem than diabetes to evaluate medical and surgical therapies in conditions such as diabetes, and emphysema.

Validity

Validity defines the range of inferences that can be made on the basis of a scale score [3]. Evidence for construct validity of health status measures is often provided by correlations with other measures of the same construct. The QWB has now been used in a wide variety of different studies. The validity evidence will be reviewed briefly and the reader is referred to the specific papers. The QWB has been used in both population and clinical studies. Figure 1 summarizes QWB scores, estimated from the US National Health Interview Survey, in relation to three problems, sinusitis, diabetes, and emphysema. In each of three age groups, sinus disease was shown to be a less serious problem than diabetes which, in turn, had less impact than emphysema [7].

Figure 2 summarizes the relationship between the QWB and measures of cognitive impairment for patients with Alzheimer’s Disease. Those with disease are significantly lower on the QWB and the degree of illness is systematically related to QWB score [11]. A variety of studies have evaluated the relationship between QWB and disease caused by the human immune deficiency virus (HIV). A series of studies was conducted at the University of California, San Diego (UCSD) HIV Neurobehavioral Research Center (HNRC). These studies evaluated the relationship between neurocognitive impairment, disability, and the QWB. Figure 3 summarizes some of the relationships between the QWB, and a variety of outcomes. When QWB scores were broken down by HIV grouping, the Centers for Disease Control (CDC) IV group was significantly lower (0.661) than the CDC A/B groups (0.755) and the control groups (0.802) (upper left panel). The differences between the class IV and class A/B is about 10 units of well-being, suggesting that individuals lose 0.1 equivalents of well years of life for each year they are in the AIDS category in comparison to the asymptomatic groups. In comparison to the uninfected controls, this would equal 1 year of life loss for each 10 infected individuals. The QWB was shown to be significantly associated with CD4+ lymphocytes (P<0.001) (upper middle), β-2 microglobulin quartile (P<0.05) (upper right), neurologists ratings of dysfunction (P<0.001) (lower left), clinician ratings neuropsychological impairment (P<0.04) (lower center), future vital status (P<0.05) (lower right) and several psychiatric variables including Profile of Mood States (POMS) scores for vigor (P<0.001) and dejection (P<0.001). Multivariate models demonstrated high covariation between predictors of QWB. These results suggest that the QWB is a significant correlate of biological, neuropsychological, neurological, psychiatric, and mortality outcomes for male HIV infected patients [12].

The QWB has also been used in clinical trials and studies to evaluate medical and surgical therapies in conditions such as chronic obstructive pulmonary disease [13], AIDS [14], cystic fibrosis [15], diabetes mellitus [16], atrial fibrillation [17], lung transplantation [18], arthritis [19], cancer [5], depression [20,21] and several other conditions [22]. Further, the method has been used for health resource allocation modeling and has served as the basis for an innovative experiment on rationing of health care by the state of Oregon [23,24]. Studies have also demonstrated that the QWB is responsive to clinical change derived from surgery [25] or medical conditions such as rheumatoid arthritis [26], AIDS [27] and cystic fibrosis [28].

QWB-SA 1.04

One of the major concerns about the QWB is that it must be administered by a trained interviewer. Although we still believe that the interviewer-administered QWB is the optimal way to collect health outcome information, we have recently developed a self-administered form, which is known as the Quality of Well-Being Self-Administered (QWB-SA) version 1.04. The self-administered form can be completed in about 10 minutes and can be machine scored. Initial data on the validity and reliability of the QWB-SA 1.04 have been reported recently [29].
In summary, the general QWB and QWB-SA 1.04 have evidence supporting validity in a variety of different specific diseases. The measures have been shown to be responsive to change and their application has been found to be feasible in population studies and clinical trials.

Comparisons between the QWB and SF-36: theory

The SF-36

The SF-36 has established itself as the most commonly used quality of life measure in the world. The SF-36 grew out of work by the RAND organization and the Medical Outcomes Study (MOS) [30]. The SF-36, includes eight health concepts: physical functioning, role-physical, bodily pain, general health perceptions, vitality, social functioning, role-emotional, and mental health. The measure has been given to literally hundreds of thousands of respondents and has an excellent record of reliability and validity [31]. In the following sections we describe some of the similarities between the QWB and SF-36. Then, we will pinpoint some of the differences between these methods.

Common history

Although not commonly recognized, the QWB and SF-36 have some common roots. Both approaches are outgrowths of early attempts to develop health status indexes in the early 1970s. The QWB evolved from a health status index originally developed by J.W. Bush and his associates at UCSD. In order to develop a population health status index, Bush et al. developed scales for mobility, physical activity and social activity [3]. These scales were based on extensive reviews of a wide variety of questionnaires used for government surveys and epidemiologic monitoring studies.

At about the same time RAND was funded to conduct the Health Insurance Experiment. The San Diego group provided their health status index for use in the study. The RAND group found the measure cumbersome, but incorporated the basic scales of mobility, physical activity and social activity into their survey instrument. They divided the social activity scale into two components of social contacts and self-care. The San Diego group had always included symptoms and problems in addition to observable function. The RAND group decided not to include specific symptom components. Further, the RAND group emphasized the use of the measures as profiles while the San Diego group evolved toward the creation of a single score. Whereas the SF-36 development excluded preference weighted scoring, QWB single-score development allowed for QWB use as a profile.

As part of the Health Insurance Experiment, the RAND group completed detailed psychometric studies of the function status questionnaire along with a variety of other measures. The next step in this programmatic research endeavor was the large scale MOS. This study continued to refine measures originally developed for the Health Insurance Experiment. Ultimately, this resulted in the current SF-36.
Although the SF-36 has evolved considerably from the original function status measures, it is also important to emphasize that many of the concepts remain similar.

The choice between measures such as the QWB and the SF-36 is difficult. In the following sections we discuss several issues that may guide these decisions. We begin with a more theoretical discussion of the differences between profile and utility based measures. Then, we critique methods for comparing measures and offer some data on correlations between measures. Finally, we address the need for separate physical and mental health measures.

**Theory: profile versus utility scoring**

A variety of generic measures assess the dimension of life quality [32]. These typically include physical functioning, emotional functioning, and some symptomatic complaints. The specific dimensions vary from measure to measure and there is substantial debate about which dimensions should be included [33].

One of the most important distinctions among measures is in how data are scaled and reported. Profile approaches report a series of scores and characterize individual respondents or groups as profiles. The focus of attention is on the components of health rather than on an overall summary. An example of one profile is shown in Figure 4. The figure shows two hypothetical profiles corresponding to treatments for headache. In this hypothetical clinical trial, patients with recurrent headaches were randomly assigned to a treatment or control group. As the figure shows, those who received the treatment scored higher on measures of role-physical, bodily pain, and physical functioning. However, they scored lower on general health perceptions, vitality, and mental health. The drug, in this case, was successful in relieving headaches but produced significant drowsiness. Ultimately, clinicians must offer some general interpretation of these profiles by applying a weighting system. They must decide if they are more concerned about physical role function or about vitality in general health perceptions. Judgment about the relative importance of various dimensions is common and typically is done implicitly, arbitrarily, and in an idiosyncratic way. Thus, physicians or patients may idiosyncratically ignore a particular test result or a particular symptom because another one is more important to them. However, the process by which relative importance is evaluated can be studied explicitly and measures of perceived relative importance can become part of the measure.

These problems in interpreting profiles for clinical decision making explain why it has been difficult to use profiles for cost-effectiveness and cost-utility analysis. Comparison of different options for the use of common resources requires overall quantification of health outcomes using a common measurement unit. To a large extent, this integrating preference or utility function is the most important feature of cost-utility analysis. Despite many attempts, popular outcome measures such as the Sickness Impact Profile (SIP), SF-36, the McMaster Health Index Questionnaire, and the Nottingham

---

**SF-36 Scale**

*Figure 4* Comparison of SF-36 profiles in hypothetical treatment of headache. PF, physical functioning; RP, role-physical; BP, bodily pain; GH, general health; VT, vitality; SF, social functioning; RE, role-emotional; MH, mental health.
Health Profile (NHP) have not been useful for cost-effectiveness analysis because they have multiple outcome dimensions. Some profiles, such as the SIP and the NHP yield a single summary score, but are not scaled on a 0 to 1.0 continuum with a clear reference to death. Thus, they are not useful for cost-effectiveness or cost-utility studies.

In addition, it is sometimes difficult to capture the total clinical picture using profile measures because most treatments have side-effects as well as benefits. A successful surgery for prostate cancer, for example, might be associated with impotence and incontinence. The major challenges are in determining what it means when someone experiences a side-effect and whether the benefits justify these side-effects. How do we determine whether or not observable side-effects are important? Should the patient who feels sleepy discontinue his or her medication? Should a patient with insulin dependent diabetes mellitus discontinue therapy because he or she develops skin problems at the injection sites? Skin problems are a nuisance, but without treatment the patient would die. Often the issue is not whether treatment causes side-effects, but how we should place these side-effects within the perspective of total health. Ultimately, we must decide whether treatment produces a net benefit or a net deficit in health status.

Often determining net benefit requires difficult trade-offs. A treatment may extend life expectancy by treating a pathophysiologic abnormality. However, the treatment may reduce quality of life or life satisfaction. Preferences and utility measures attempt to quantify quality of life in relation to life expectancy. These methods can be used to estimate net benefit. The assessment of net benefit requires summary measures that scale relative desirability of various health outcomes. The scaling represents utility or preferences for components of health.

Measuring preference evokes many technical and methodological challenges [6,34]. Different methods of preference measurement can yield different results [35], a finding that should not be surprising because the various approaches to preference assessment are based on different underlying conceptual models and the methods ask different questions. Decision theory based methods are similar in that they place wellness on a continuum between 0.0 and 1.0.

Decision theory methods are refinements of generic survival analysis. In traditional survival analysis, those who are alive are statistically coded as 1.0 while those who are dead are statistically coded as 0.0. Mortality can result from any physiologic abnormality. However, treatment may reduce quality of life or life satisfaction. Preferences and utility measures attempt to quantify quality of life in relation to life expectancy. These methods can be used to estimate net benefit. The assessment of net benefit requires summary measures that scale relative desirability of various health outcomes. The scaling represents utility or preferences for components of health.

Figure 5 QWB subscale adjustments in late life psychosis patients of different levels of severity. (The data in the figure are from [37].)

Measured utilities are needed to calculate quality-adjusted life-years (QALYs). These methods are required in order to perform cost-utility analysis [36]. QALYs integrate mortality and morbidity to express health status in terms of equivalents of well-years of life. Suppose a woman who has a life expectancy of 75 years dies of breast cancer at age 50; the disease was associated with 25 lost life-years. If 100 women died at age 50 (and also had a life expectancy of 75 years) 2500 (100 x 25 years) life-years would be lost.

Death is not the only outcome of concern in cancer. Many adults suffer from the disease leaving them somewhat disabled over long periods of time. Although still alive, the quality of their lives has diminished. QALYs take into consideration the quality of life consequences of these illnesses. For example, a disease that reduces quality of life by one half will take away 0.5 QALYs over the course of 1 year. If it affects two people, it will take away 1 year (2 x 0.5) over a 1-year period. A pharmaceutical treatment that improves quality of life by 0.2 for each of five individuals will result in the equivalent of one QALY if the benefit is maintained over a 1-year period. This system has the advantage of considering both benefits and side-effects of programs in terms of the common QALY units. Although QALYs are typically assessed for patients, they can also be measured for others, including care givers who are placed at risk because they experience excess stress and burden.

One of the important contrasts between the QWB and the SF-36 is that the QWB is primarily a preference-based measure designed to produce estimates of QALYs. The SF-36 is primarily a profile measure. However, summary scores are available for the SF-36 mental and physical health components. Similarly, profile information can be obtained from the QWB and QWB-SA. In fact, some investigators prefer to report profile information along with the overall score. For example, Patterson et al. [37] evaluated QWB outcomes for patients with late onset schizophrenia. The study is interesting because it is always assumed that the variation in outcomes for patients with mental illness will be reflected by mental health items. Figure 5 summarizes some profile information from the Patterson study. It compares patients who are normal, or have low, moderate, or high degrees of dysfunction associated with schizophrenia (evaluated by formal psychiatric interviews) in relation to QWB scales for...
Table 2  Principal concepts and domains of health-related quality of life contained in general preference weighted instruments for assessing quality-adjusted life years

<table>
<thead>
<tr>
<th>Concept</th>
<th>Disability Distress Index</th>
<th>EuroQol 15D</th>
<th>Mark I</th>
<th>Mark II</th>
<th>Mark III</th>
<th>Years of Healthy Life</th>
<th>Quality of Well-being</th>
<th>Quality of Life and Health</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health perceptions</td>
<td></td>
<td></td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>Social function</td>
<td></td>
<td></td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social relations</td>
<td></td>
<td></td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Usual social role</td>
<td></td>
<td></td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intimacy/sexual function</td>
<td></td>
<td></td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Communication/speech</td>
<td></td>
<td></td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Psychological function</td>
<td></td>
<td></td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cognitive function</td>
<td></td>
<td></td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
<td></td>
<td>**</td>
</tr>
<tr>
<td>Emotional function</td>
<td></td>
<td></td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>Mood/feelings</td>
<td></td>
<td></td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>Physical function</td>
<td></td>
<td></td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>Mobility</td>
<td></td>
<td></td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>Physical activity</td>
<td></td>
<td></td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>Self-care</td>
<td></td>
<td></td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>Impairment</td>
<td></td>
<td></td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensory function/loss</td>
<td></td>
<td></td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Symptoms/impairments</td>
<td></td>
<td></td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


mobility, physical activity, and social activity. The scaling was the adjustment (from 1.0) in the QWB scale. Thus, higher scores in the figure represent more dysfunction. As the figure demonstrates, those with more severe schizophrenia have higher adjustments on the QWB subscales. The figure is interesting because it demonstrates that QWB profile information can be reported if the investigator desires. Further, the figure demonstrates that traditional functional scales capture variability associated with mental illness. We will review the issue of mental health in more detail in the last section of the paper.

**Comparisons between the QWB and SF-36: studies**

**Content and face validity**

In 1993, the US Department of Health and Human Services appointed a multi-disciplinary group of methodologists to recommend standardized strategies for the evaluation of health care. The panel, which released its report in 1996, suggested that standardized outcomes analyses be conducted to evaluate the cost-effectiveness of medical care [36]. These analyses require preference weighted measures of health related quality of life. Although there has been considerable interest in measuring the cost-effectiveness of treatments, little is known about the validity of general outcome measures and it is often difficult to choose between different approaches. Some authors have attempted to simplify the task by offering summary tables. An example of one such summary is shown as Table 2.

Despite the attractiveness of this approach, there are also some difficulties. In particular, creators of the tables typically examine the names of subscales, rather than the content of the measures. Consider the example of sensory function or loss. According to Table 2, sensory functioning is not included in the QWB. The creators of the table came to this conclusion because there is no subscale on the QWB named sensory function. However, the QWB-SA includes symptoms for loss of vision, loss of hearing, impairment of vision (including wearing glasses or contact lenses), problems with taste and smell, and so on. In fact the newer self-administered QWB (QWB-SA) devotes a major portion of the questionnaire to items on sensation and sensory organs. The symptoms include any hearing loss, blindness in one eye, blindness in both eyes, any problems with vision (floaters, double or distorted vision), eye pain, sensitivity to light, ear aches, difficulty in balance, and a variety of others. Indeed, the QWB-SA includes much more content on sensory functioning than do measures that are identified as including content on sensory functioning.
Correlations between the QWB and SF-36

A growing number of studies have applied both the SF-36 and the QWB. Table 3 summarizes correlations between the QWB and SF-36 components in four recent studies. The first study is based on a population sample in Beaver Dam, Wisconsin involving 1356 adults (560 men and 796 women). The mean age of the sample was 64.1 years. In this sample, all of the participants lived in community settings. The second study involved 201 adults with serious illnesses, including 99 with AIDS, 74 with cancer, and 28 with other terminal illnesses. The AIDS patients were younger (mean = 38 years) while those with cancer were older (mean = 61 years). Those with other illnesses were also older (mean = 65 years). The third sample was a group of 100 HIV infected men (mean age 35.3 years). These men completed the MOS-HIV-36 which is a 34-item adaptation of the SF-36 for people with HIV disease. The fourth sample included 301 community dwelling older adults (124 men and 177 women). All participants were older than 65 years and 44% of the men and 53% of the women were older than 75 years.

In all studies correlations between the QWB total score and the SF-36 physical function scale are substantial. Similarly, correlations between the QWB scale and the role-emotional scale are low in all four studies. As the table shows, correlations between the QWB and SF-36 components are quite consistent across these very different patient populations. The only exception is the low correlation between total QWB and bodily pain in the terminally ill sample. Overall, the available evidence suggests that the SF-36 and QWB tap most of the same variation in health status.

Ceiling effects

Measures differ in the extent to which they are responsive to minor variations in wellness. Some of our early studies demonstrated that nearly 80% of the general population has no functional limitations on a particular day. Thus, functional items on surveys such as the National Health Interview Survey show the great majority of people to be well. On the other hand, only about 12% of the general population experience no symptom or problem during an average week. Fryback et al. [38] studied the distributional properties of the QWB and SF-36 scales in Beaver Dam, Wisconsin. They found the QWB to be approximately normally distributed with scores of 1.0 being exceptionally rare. This finding has also been reported by our group [3] and others.

Ganiats et al. [42] have evaluated the QWB, SF-36, and other measures in clinical trials for patients with atrial fibrillation. Confirming other studies, they found 0% of patients at the ceiling level for the QWB. In contrast, many SF-36 scales were at their ceiling level. In other words, improvements on these dimensions would be difficult to detect because the highest values had already been recorded. For example, nearly 70% of respondents obtained the highest score for role-emotional and 54% obtained the highest score for social functioning. This was a concern because these patients were afflicted with a serious chronic illness. However, the SF-36 scores for this elderly population were comparable to those from the MOS normative sample (Figure 6).

One of the issues in the application of QWB, SF-36, and other measures is the percentage of respondents who have incomplete questionnaires. In the atrial fibrillation studies, all measures were administered to at least 300 patients. The percentage of patients who had any data missing for each measure was recorded. Because it was interviewer administered, no patient had incomplete survey responses for the QWB. The EQ-5D had 6% incomplete survey responses. The Health Utility Index (HUI) had 11% and the SF-36 had 26%. The new self-administered QWB also has had some missing data problems similar to those of other measures. In the recent study by Andersen et al. [41], missing data for the QWB-SA were common: about 3% of the respondents failed to fill out items on chronic diseases. Overall, the missing
data rate was about the same as it is for the SF-36. However, the SF-36 has worked out a scoring routine that allows the measures to be scored with up to 50% of the items missing. We are currently working on a missing data scoring protocol for the QWB-SA.

Mental health

Despite widespread interest in the concept of a QALY among practitioners in many different specialties, these ideas have received very little attention in the mental health fields. This reflects the widespread belief that mental health and physical health outcomes are conceptually distinct.

Although many questionnaires include different dimensions, they still may be tapping the same constructs. For example, a measure without a mental health component does not necessarily neglect mental health. Mental health symptoms may be included and the impact of mental health, cognitive functioning, or mental retardation may be represented in questions about role functioning. Some measures have multiple dimensions for mental health symptoms while others include fewer items that ask about problems in general. It is not clear that multiple measures are more capable of detecting clinical differences. This remains an empirical question for systematic analysis. A common strategy is to report outcomes along multiple dimensions.

Several years ago Kaplan and Anderson [43] argued that there are many similarities in mental health and physical health outcomes. The preference- and utility-based measures, that are assumed to ignore mental health, include the basic dimensions of observable functioning, symptoms, and duration. Mental health problems, like physical health problems, can be represented by symptoms and by disrupted role functioning. Consider some examples. Suppose that a patient has the primary symptom of a cough. If the cough does not disrupt role function, the preference or utility weighted score might show a small deviation from 1.0. If the cough is more serious and keeps the person at home, the score will be lower. If the cough is very severe, it might limit the person to a hospital and may have serious disruptive effects upon role functioning. This would necessitate an even lower score. Coughs can be of different duration. A cough associated with an acute respiratory infection may have a serious impact on functioning that may last only a short period of time. This would be indicated by a minor deviation in QALY's. A chronic cough associated with obstructive lung disease would be associated with significant loss of QALY's because duration is a major component of the calculation.

Now consider the case of a person with depression. Depression may be a symptom reported by a patient just as a cough is reported by other patients. Depression without disruption of role function would cause a minor variation of wellness. If the depression caused the person to stay at home the preference or utility weighted score would be lower. Severe depression might require the person to be in a hospital or special facility and would result in a lower score. Depressions, like coughs, are of different durations. Depression of long duration would cause the loss of more QALY's than would depression of short duration.

It is commonly asserted that the QWB excludes mental health content. Existing empirical evidence supports the validity of the QWB in studies of patients known to have impaired mental health. One study evaluated the validity of the QWB as an outcome measure for older psychotic patients. Seventy-two psychotic patients and 28 matched controls from the San Diego Veterans Affairs Medical Center completed the QWB; the Structured Clinical Interview for the DSM-III-R patient version, Scales for the Assessment of Positive and Negative Symptoms (SAPS and SANS), and the Global Severity Index (GSI) from the Brief Symptom Inventory were administered to all subjects. The QWB was significantly correlated with the SANS \( r = -0.52, P < 0.001 \) the SAPS \( r = -0.57, P < 0.001 \) and the GSI \( r = -0.62, P < 0.001 \). There was a strong and significant linear relationship between QWB and severity of illness (as classified by the SANS and the SAPS). In addition, component scores of the QWB (i.e., mobility, physical activity, social activity, and worst symptom) were significantly lower among patients as compared to controls, and declined systematically as psychiatric symptoms increased [37]. In a related study, Rapaport et al. [44] found that patients with schizophrenia had QWB scores that were significantly lower \( (0.6) \) than age matched non-schizophrenia controls \( (0.7) \). QWB differences between these two groups were comparable in magnitude to a wide array of psychiatry specific measures.

Several other studies have now documented the sensitivity of the QWB to mental health problems. For example, a variety of studies has demonstrated the validity of the QWB for assessing depression among patients with HIV disease. In one study, ratings of depression using the Hamilton Depression Scale (HAM-D) were obtained from 285 HIV patients and 84 HIV negative men participating in the HNRC cohort. The data were obtained at baseline and 6 months later. Depression was defined as Hamilton scores greater than 10. The study demonstrated a systematic relationship between Hamilton scores and QWB scores at baseline [45]. In addition, 22 HIV-positive subjects experienced increases of 10 points on the Hamilton Scale between the first evaluation and that at 6 months. For these individuals, significant reductions in QWB scores were observed. Analyses of QWB symptoms suggested greater symptom severity among those whose Hamilton scores increased. In addition, there was a greater reduction in physical activity. In other words, mood affected both symptoms and physical function.

Pyne et al. [20,21] compared QWB scores between patients with major depression and controls. The depressed patients were divided into mild, moderate and severe groups according to their scores on the HAM-D. There was a step-wise decrease in QWB scores for the controls through each of the patient groups. Similar results were observed for Beck Depression Inventory (BDI) scores. The difference between each of these groups is highly significant \( (P<0.001) \). In order to evaluate which of six variables (age, sex, family history, presence of Axis III diagnosis, comorbid Axis I diagnosis, HAM-D) best predicted the QWB score in this patient population, a hierarchical multiple regression analysis was
conducted. Although Axis III disorders are significant predictors of QWB, depression (HAM-D and BDI) predicts QWB when Axis III diagnoses are controlled statistically.

The Beaver Dam study offers some evidence that the QWB is sensitive to mental health disorders in the community. The study offers QWB scores for individuals with a wide variety of conditions. Respondents with self-reported depression had the third lowest mean QWB score among all conditions [46].

QALYs can be used to compare treatments for physical and mental health problems. Suppose, for example, that a treatment for anxiety elevates patients from a level of 0.65 to a level of 0.75. Suppose further, that this treatment benefit lasted for 1 year. Each patient would gain 0.10 QALY (0.75−0.65 = 0.10 × 1 year = 0.10 QALY) for each year the benefit was observed. The treatment benefit would be expressed in terms of general QALY units. The productivity of the providers could be compared with providers in other areas of health care. All providers in health care use resources. Dividing the cost of a treatment by the QALY productivity provides the cost-utility ratio. Measuring productivity of mental health-related treatments in QALY units would allow the investments in mental health services to be compared directly to those in other aspects of health care.

One of the challenges to single-score measures is the argument that physical and mental health are different dimensions and that any measure combining them into a single index is like combining apples and oranges (CE Schwartz, RM Kaplan, JP Anderson, T Holbrook and MW Genderson, unpublished work). Clearly, mental and physical health services are different. Yet, providers compete for the same health care resources and there must be some basis for deciding how to allocate these resources.

We have conducted several studies designed to determine if mental and physical health clearly separate into well-defined dimensions. Ware et al. [31] have clearly shown that factor analyses of the SF-36 suggest separate physical and mental health dimensions. In a series of studies, we factor analyzed the QWB using groups of patients with multiple sclerosis (n = 263), non-insulin dependent diabetes mellitus (n = 420), non-head traumatic injury (n = 852), AIDS (n = 99), and cancer (n = 74). Each QWB item was treated as dichotomous and the matrix of 0s and 1s was subjected to factor analysis with orthogonal rotation. Items with factor loadings > 0.40 were regarded as consistent with a particular factor.

The analysis consistently revealed two strong factors. The first factor combined physical and psychological symptoms. The same factor emerged across these diverse patient populations. The items that consistently load on this factor include upset stomach, general tiredness, feeling upset, depressed or crying, dizziness, trouble sleeping, and excessive worry. In contrast to the symptom factor, the second dimension was defined by function. The second factor includes items such as spending the day in a wheelchair, difficulty in walking, and limitations in social activity (CE Schwartz, RM Kaplan, JP Anderson, T Holbrook and MW Genderson, unpublished work).

In three of these four populations, SF-36 data were also available. Using the same strategy, we have replicated factor analyses showing that mental and physical health separation for the SF-36. How might we explain why the QWB factor analysis mixes mental and physical health symptoms, whereas the SF-36 factor analysis separates them? One explanation is the difference between symptoms and functioning. In both factor analyses, functioning items emerge as a separate factor. The mental health items on the SF-36 use a different format and place greater emphasis on symptomatic activity. The SF-36 also separates instructions for role-emotional and physical functioning scales. Respondents are instructed to consider disruptions in activities that result from physical or mental health problems. Thus, separation of physical and emotional components in factor analysis might be expected because respondents had been prompted to think differently about them. It is worth noting that the separate dimensions of physical and mental health reported by Ware et al. [31] in part from their use of orthogonal rotations techniques. Methods that accommodate correlated factors (i.e. oblique rotation) have been shown to offer a better fit for health data [47].

Others have noted the complications resulting from the separation of physical and mental health. For example, Simon et al. evaluated 536 primary care patients before and after treatment with antidepressant medications. The mental health treatment was associated with improvements in both mental and physical health subscales of the SF-36 (physical function, role-physical, bodily pain, and general health perceptions). However, because of an artifact in the scoring system, the physical health summary score was unchanged. The study is important for two reasons: it demonstrates the complications in interpreting SF-36 summary scores; and it demonstrates that mental health treatments may have significant effects on measures believed to represent physical health [48]. We believe this underscores the fuzzy boundary between physical and mental health.

Summary

Outcomes researchers now have a variety of validated approaches to assess the costs, risks and benefits in medical care. The QWB and the SF-36 are two methods that arise from similar traditions. Many components of the measures are similar and correlations between the measures are substantial. Until recently, the SF-36 was much easier and less expensive to administer. However, an inexpensive self-administered form of the QWB is now available.

There are also important distinctions between the QWB and SF-36. The QWB places greater emphasis on symptoms and provides more clinical information. For example, it offers the clinician symptoms reports similar to a review of systems. Further, the QWB can be used for policy analysis because outcomes can be translated into QALYs. Although it is possible to create profiles from the QWB, it does less well than the SF-36 for characterizing multi-dimensional patterns of outcome. The QWB and SF-36 represent different
measurement approaches. The SF-36 is rooted in psychometric theory. The QWB, although adhering to some principles of psychometric theory, arises from a decision theory tradition. Because of the way it is constructed, some aspects of psychometric theory do not apply directly to the QWB. For example, traditional test–retest reliability has little value for assessing either the QWB or the SF-36. The QWB is better suited to policy analysis and to economic studies that require the calculation of a QALY. Investigators interested in cost-effectiveness or cost-utility analysis should consider using the QWB or a related utility-based measure. Investigators interested in reviewing a profile of outcomes may be better to use the well-established and well-validated SF-36.

We believe that better measurement technologies will replace many current approaches to health status assessment. Continuing research is necessary to build the next-generation of outcome measures.

Acknowledgements

This work was supported in part by grants R01 HS 09170 and 5RO1NS33519 from the Agency for Health Care Policy and Research and grant P60 AR 40770 from the National Institutes of Health.

References

519


Accepted for publication 4 September 1998
The Quality of Well-Being Scale: critical similarities and differences with SF-36

ROBERT M. KAPLAN, THEODORE G. GANIATS, WILLIAM J. SIEBER AND JOHN P. ANDERSON

Department of Family and Preventive Medicine, University of California, CA, USA

Abstract

Purpose. To summarize the development and application of a generic measure of health-related quality of life known as the Quality of Well-Being Scale (QWB).

Background. The QWB is part of a general health policy model. The measure includes functional components for mobility, physical activity, and social activity. In addition, it includes a comprehensive list of symptoms and problems. QWB scoring allows placement of each individual on a continuum of wellness ranging from 0 (for dead) to 1.0 for asymptomatic full function. The General Health Policy Model combines this point in time measure with information on prognosis and mortality to estimate quality-adjusted life-years (QALYs). Forming a ratio of program costs to QALYs yields estimates of cost per quality-adjusted life year. Evidence supports the validity of the QWB for a wide variety of applications in population monitoring, descriptive studies of patient populations, and clinical trials. We offer a variety of comparisons between the QWB and Medical Outcomes Study Short Form 36 (SF-36).

Main findings. In comparison with the SF-36, the QWB provides less information on health profiles, but has the advantage of providing a metric that can be used for cost-utility or cost-effectiveness analysis. Although ceiling effects are common for some SF-36 subscales, perfect scores on the QWB are very rare. The QWB has an approximately normal distribution for populations of adults. Although often criticized for not including a mental health component, we present evidence documenting the validity of the QWB for patients with psychological and psychiatric diagnoses.

Conclusions. The QWB and SF-36 are alternative comprehensive measures of health outcomes.

Keywords: outcomes assessment, outcomes research, quality of life, Quality of Well-Being Scale (QWB), SF-36

A variety of different methods are now available to measure health-related quality of life. Many of these measures are specific to particular illnesses or diseases. However, there is a continuing need for general or generic measures that can be used for population monitoring, evaluation research, individual clinical decisions, or as outcome measures in randomized clinical trials. The purposes of this paper are to: (i) review the Quality of Well-Being Scale (QWB) and (ii) compare the QWB with the Medical Outcomes Study 36 Item Short Form (SF-36).

QWB–general health policy model

General background

The General Health Policy Model grew out of substantive theories in economics, psychology, medicine, and public health. The model of health status includes components for mortality (death) morbidity (health-related quality of life) and time. The rationale for the model is that diseases and disabilities are important for two reasons. First, illness may cause the life expectancy to be shortened. Second, illness may make life less desirable at times prior to death (health-related quality of life) [1–5].

Central to the General Health Policy Model is a general conceptualization of quality of life. The QWB is a method of measuring quality of life for calculations in the model. The QWB is a preference-weighted measure combining three scales of functioning with a measure of symptoms and problems to produce a point-in-time expression of well-being that runs from 0 (for death) to 1.0 (for asymptomatic full function) [3]. The model separates aspects of health status and life quality into distinct components. These are life expectancy (mortality), functioning and symptoms (morbidity), preference for observed functional states (utility) and


Address correspondence to Robert M. Kaplan, University of California, San Diego, Department of Family and Preventative Medicine, 9500 Gilman Drive, 0628 La Jolla, CA 92093–0628, USA. Tel: +1 619 534 6058. Fax: +1 619 534 7517. E-mail: rkaplan@ucsd.edu

© 1998 International Society for Quality in Health Care and Oxford University Press
Reliability

Reliability is an estimate of the proportion of variance in a test or measure that is true score. Subtracting this proportion from 1.0 gives an estimate of the proportion of variation assumed to be error. There are several potential sources of error. Two important sources are item sampling and time sampling. One of the basic tenets of psychometric theory is that each item in a test or measure is an unbiased and representative sample from the domain under study. In the construction of an intelligence test, for example, there is an infinite number of items that might represent intellectual ability. The Domain Sampling Model [9] assumes that items are sampled from this domain of performance and that each item is assumed to be an unbiased estimate of the underlying trait. Reliability is estimated from the inter-item correlations and the extent to which the items are intercorrelated characterizes the reliability of the measure. Reliable measures are those for which the component items measure the same construct.

Measures of health status may be derived from a very different theoretical model. Items may not be considered random samples from a large domain because each may have a very specific meaning. For example, report of a severe headache offers very specific information. The item is not randomly sampled from all possible symptoms and the meaning of reporting a headache is very different from the meaning of reporting difficulty urinating. Thus, item sampling, as known in psychometric theory, is not necessarily relevant to many health status measures. A question about a headache, as used in clinical medicine, is not randomly selected from a large domain of questions about health.

A second type of reliability in psychometric theory concerns time sampling. Psychological traits are considered to be stable over the course of time. If Sally is intelligent today, we expect her to be equally intelligent 2 weeks from today. Variation in her performance across the 2 weeks might be attributable to measurement error. Thus, test-retest estimates are an important source of information.

This same logic may not apply to health status measures. If Sally is very sick today, we may not expect her to be equally sick in 2 weeks time. Differences between health scores taken at two points in time may mean that she recovered from her illness or that she got sicker. When the underlying construct is expected to change over time, test-retest evaluations may have very little meaning. As a result, traditional reliability data have less meaning for the QWB.

We are not implying that utility measures are exempt from reliability assessment. Measurement error is assessed in several other ways. For example, we consider reliability assessed over short intervals when large swings in health status are not expected. In addition, we can consider agreement on classification of functioning using different methods.

Considerable evidence shows that the QWB scores assessed on consecutive days are highly reliable. Table 1 summarizes consecutive day QWB correlations for a variety of populations. As the table shows, the measure has good short term stability [10]. A second method for calculating interday reliability estimated agreement per cent for reports of dysfunction on consecutive days. The agreement percentage (AP) is calculated as:

\[ AP = \frac{\text{no. of agreements}}{\text{no. of agreements} + \text{no. of disagreements}} \]

These values are shown in the righthand column of Table 1. The values were not available for three of the seven populations. The reported values are from Anderson et al. [10].

<table>
<thead>
<tr>
<th>Population</th>
<th>Reliability</th>
<th>Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>General population adults ((n = 681))</td>
<td>0.96</td>
<td>0.82</td>
</tr>
<tr>
<td>General population children ((n = 274))</td>
<td>0.93</td>
<td>0.87</td>
</tr>
<tr>
<td>Adults in California—Indochinese community ((n = 598))</td>
<td>0.94</td>
<td>0.94</td>
</tr>
<tr>
<td>Burn patients ((n = 143))</td>
<td>0.83</td>
<td>0.97</td>
</tr>
<tr>
<td>Non-head trauma patients ((n = 1048))</td>
<td>0.90</td>
<td></td>
</tr>
<tr>
<td>Chronic obstructive pulmonary disease patients ((n = 84))</td>
<td>0.98</td>
<td></td>
</tr>
<tr>
<td>Diabetes patients ((n = 70))</td>
<td>0.96</td>
<td></td>
</tr>
</tbody>
</table>

1Data from [10] and JP Anderson, unpublished work.
2Agreement data not available for trauma, COPD and diabetes patients.
QWB versus SF-36

Figure 1  Sinus disease and diabetes in the general population. (The data in the figure are from [7].)

Validity

Validity defines the range of inferences that can be made on the basis of a scale score [3]. Evidence for construct validity of health status measures is often provided by correlations with other measures of the same construct. The QWB has now been used in a wide variety of different studies. The validity evidence will be reviewed briefly and the reader is referred to the specific papers. The QWB has been used in both population and clinical studies. Figure 1 summarizes QWB scores, estimated from the US National Health Interview Survey, in relation to three problems, sinusitis, diabetes, and emphysema. In each of three age groups, sinus disease was shown to be a less serious problem than diabetes which, in turn, had less impact than emphysema [7].

Figure 2 summarizes the relationship between the QWB and measures of cognitive impairment for patients with Alzheimer’s Disease. Those with disease are significantly lower on the QWB and the degree of illness is systematically related to QWB score [11]. A variety of studies have evaluated the relationship between QWB and disease caused by the human immune deficiency virus (HIV). A series of studies was conducted at the University of California, San Diego (UCSD) HIV Neurobehavioral Research Center (HNRC). These studies evaluated the relationship between neurocognitive impairment, disability, and the QWB. Figure 3 summarizes some of the relationships between the QWB, and a variety of outcomes. When QWB scores were broken down by HIV grouping, the Centers for Disease Control (CDC) IV group was significantly lower (0.661) than the CDC A/B groups (0.755) and the control groups (0.802) (upper left panel). The differences between the class IV and class A/B is about 10 units of well-being, suggesting that individuals lose 0.1 equivalents of well years of life for each year they are in the AIDS category in comparison to the asymptomatic groups. In comparison to the uninfected controls, this would equal 1 year of life loss for each 10 infected individuals. The QWB was shown to be significantly associated with CD4 + lymphocytes (P<0.001) (upper middle), β-2 microglobulin quartile (P<0.05) (upper right), neurologists ratings of dysfunction (P<0.001) (lower left), clinician ratings neuropsychological impairment (P<0.04) (lower center), future vital status (P<0.05) (lower right) and several psychiatric variables including Profile of Mood States (POMS) scores for vigor (P<0.001) and dejection (P<0.001). Multivariate models demonstrated high covariation between predictors of QWB. These results suggest that the QWB is a significant correlate of biological, neuropsychological, neurological, psychiatric, and mortality outcomes for male HIV infected patients [12].

The QWB has also been used in clinical trials and studies to evaluate medical and surgical therapies in conditions such as chronic obstructive pulmonary disease [13], AIDS [14], cystic fibrosis [15], diabetes mellitus [16], atrial fibrillation [17], lung transplantation [18], arthritis [19], cancer [5], depression [20,21] and several other conditions [22]. Further, the method has been used for health resource allocation modeling and has served as the basis for an innovative experiment on rationing of health care by the state of Oregon [23,24]. Studies have also demonstrated that the QWB is responsive to clinical change derived from surgery [25] or medical conditions such as rheumatoid arthritis [26], AIDS [27] and cystic fibrosis [28].

QWB-SA 1.04

One of the major concerns about the QWB is that it must be administered by a trained interviewer. Although we still believe that the interviewer-administered QWB is the optimal way to collect health outcome information, we have recently developed a self-administered form, which is known as the Quality of Well-Being Self-Administered (QWB-SA) version 1.04. The self-administered form can be completed in about 10 minutes and can be machine scored. Initial data on the validity and reliability of the QWB-SA 1.04 have been reported recently [29].
In summary, the general QWB and QWB-SA l.04 have evidence supporting validity in a variety of different specific diseases. The measures have been shown to be responsive to change and their application has been found to be feasible in population studies and clinical trials.

Comparisons between the QWB and SF-36: theory

The SF-36

The SF-36 has established itself as the most commonly used quality of life measure in the world. The SF-36 grew out of work by the RAND organization and the Medical Outcomes Study (MOS) [30]. The SF-36, includes eight health concepts: physical functioning, role-physical, bodily pain, general health perceptions, vitality, social functioning, role-emotional, and mental health. The measure has been given to literally hundreds of thousands of respondents and has an excellent record of reliability and validity [31]. In the following sections we describe some of the similarities between the QWB and SF-36. Then, we will pinpoint some of the differences between these methods.

Common history

Although not commonly recognized, the QWB and SF-36 have some common roots. Both approaches are outgrowths of early attempts to develop health status indexes in the early 1970s. The QWB evolved from a health status index originally developed by J.W. Bush and his associates at UCSD. In order to develop a population health status index, Bush et al. developed scales for mobility, physical activity and social activity [3]. These scales were based on extensive reviews of a wide variety of questionnaires used for government surveys and epidemiologic monitoring studies.

At about the same time RAND was funded to conduct the Health Insurance Experiment. The San Diego group provided their health status index for use in the study. The RAND group found the measure cumbersome, but incorporated the basic scales of mobility, physical activity and social activity into their survey instrument. They divided the social activity scale into two components of social contacts and self-care. The San Diego group had always included symptoms and problems in addition to observable function. The RAND group decided not to include specific symptom components. Further, the RAND group emphasized the use of the measures as profiles while the San Diego group evolved toward the creation of a single score. Whereas the SF-36 development excluded preference weighted scoring, QWB single-score development allowed for QWB use as a profile. As part of the Health Insurance Experiment, the RAND group completed detailed psychometric studies of the function status questionnaire along with a variety of other measures. The next step in this programmatic research endeavor was the large scale MOS. This study continued to refine measures originally developed for the Health Insurance Experiment. Ultimately, this resulted in the current SF-36.

Figure 3 Relationship between QWB and CDC group (A), CD4+ cells (B), b 2-microglobulin (C), neurologists ratings of central nervous system dysfunction (D), neuropsychological test summary score of impairment (E), and eventual death within 18 months (F). Error bars show 95% confidence intervals. (The data in the figure are from [14]).
The figure shows two hypothetical profiles corresponding to cost-utility analysis. The specific dimensions vary from measure to measure, and there is substantial debate about which dimensions should be included [33].

Although the SF-36 has evolved considerably from the original function status measures, it is also important to emphasize that many of the concepts remain similar.

The choice between measures such as the QWB and the SF-36 is difficult. In the following sections we discuss several issues that may guide these decisions. We begin with a more theoretical discussion of the differences between profile and utility based measures. Then, we critique methods for comparing measures and offer some data on correlations between measures. Finally, we address the need for separate physical and mental health measures.

**Theory: profile versus utility scoring**

A variety of generic measures assess the dimension of life quality [32]. These typically include physical functioning, emotional functioning, and some symptomatic complaints. The specific dimensions vary from measure to measure and there is substantial debate about which dimensions should be included [33].

One of the most important distinctions among measures is in how data are scaled and reported. Profile approaches report a series of scores and characterize individual respondents or groups as profiles. The focus of attention is on the components of health rather than on an overall summary. An example of one profile is shown in Figure 4. The figure shows two hypothetical profiles corresponding to treatments for headache. In this hypothetical clinical trial, patients with recurrent headaches were randomly assigned to a treatment or control group. As the figure shows, those who received the treatment scored higher on measures of role-physical, bodily pain, and physical functioning. However, they scored lower on general health perceptions, vitality, and mental health. The drug, in this case, was successful in relieving headaches but produced significant drowsiness. Ultimately, clinicians must offer some general interpretation of these profiles by applying a weighting system. They must decide if they are more concerned about physical role function or about vitality in general health perceptions. Judgment about the relative importance of various dimensions is common and typically is done implicitly, arbitrarily, and in an idiosyncratic way. Thus, physicians or patients may idiosyncratically ignore a particular test result or a particular symptom because another one is more important to them. However, the process by which relative importance is evaluated can be studied explicitly and measures of perceived relative importance can become part of the measure.

These problems in interpreting profiles for clinical decision making explain why it has been difficult to use profiles for cost-effectiveness and cost-utility analysis. Comparison of different options for the use of common resources requires overall quantification of health outcomes using a common measurement unit. To a large extent, this integrating preference or utility function is the most important feature of cost-utility analysis. Despite many attempts, popular outcome measures such as the Sickness Impact Profile (SIP), SF-36, the McMaster Health Index Questionnaire, and the Nottingham

![SF-36 Scale](image-url)

**Figure 4** Comparison of SF-36 profiles in hypothetical treatment of headache. PF, physical functioning; RP, role-physical; BP, bodily pain; GH, general health; VT, vitality; SF, social functioning; RE, role-emotional; MH, mental health.
Health Profile (NHP) have not been useful for cost-effectiveness analysis because they have multiple outcome dimensions. Some profiles, such as the SIP and the NHP yield a single summary score, but are not scaled on a 0 to 1.0 continuum with a clear reference to death. Thus, they are not useful for cost-effectiveness or cost-utility studies.

In addition, it is sometimes difficult to capture the total clinical picture using profile measures because most treatments have side-effects as well as benefits. A successful surgery for prostate cancer, for example, might be associated with impotence and incontinence. The major challenges are in determining what it means when someone experiences a side-effect and whether the benefits justify these side-effects. How do we determine whether or not observable side-effects are important? Should the patient who feels sleepy discontinue his or her medication? Should a patient with insulin dependent diabetes mellitus discontinue therapy because he or she develops skin problems at the injection sites? Skin problems are a nuisance, but without treatment the patient would die. Often the issue is not whether treatment causes side-effects, but how we should place these side-effects within the perspective of total health. Ultimately, we must decide whether treatment produces a net benefit or a net deficit in health status.

Often determining net benefit requires difficult trade-offs. A treatment may extend life expectancy by treating a pathophysiologic abnormality. However, the treatment may reduce quality of life or life satisfaction. Preferences and utility measures attempt to quantify quality of life in relation to life expectancy. These methods can be used to estimate net benefit. The assessment of net benefit requires summary measures that scale relative desirability of various health outcomes. The scaling represents utility or preferences for components of health.

Measuring preference evokes many technical and methodological challenges [6,34]. Different methods of preference measurement can yield different results [35], a finding that should not be surprising because the various approaches to preference assessment are based on different underlying conceptual models and the methods ask different questions. Decision theory based methods are similar in that they place wellness on a continuum between 0.0 and 1.0.

Decision theory methods are refinements of generic survival analysis. In traditional survival analysis, those who are alive are statistically coded as 1.0 while those who are dead are statistically coded as 0.0. Mortality can result from any disease and survival analysis allows the comparison between different diseases. For example, we can state the life expectancy for those who will eventually die of heart disease and compare it to the life expectancy to those who eventually die of cancer. Thus, there is an advantage over disease specific measures such as heart ejection fractions and tumor size. The difficulty is that everyone who remains alive is given the same score. A person confined to bed with an irreversible coma is alive and is counted the same as someone who is actively participating in athletics. Utility assessment, on the other hand, allows the quantification of levels of wellness on the continuum anchored by death and optimum function.

![Figure 5 QWB subscale adjustments in late life psychosis patients of different levels of severity. (The data in the figure are from [37].)](image)

Measured utilities are needed to calculate quality-adjusted life-years (QALYs). These methods are required in order to perform cost-utility analysis [36]. QALYs integrate mortality and morbidity to express health status in terms of equivalents of well-years of life. Suppose a woman who has a life expectancy of 75 years dies of breast cancer at age 50; the disease was associated with 25 lost life-years. If 100 women died at age 50 (and also had a life expectancy of 75 years) 2500 (100 x 25 years) life-years would be lost.

Death is not the only outcome of concern in cancer. Many adults suffer from the disease leaving them somewhat disabled over long periods of time. Although still alive, the quality of their lives has diminished. QALYs take into consideration the quality of life consequences of these illnesses. For example, a disease that reduces quality of life by half will take away 0.5 QALYs over the course of 1 year. If it effects two people, it will take away 1 year (2 x 0.5) over a 1-year period. A pharmaceutical treatment that improves quality of life by 0.2 for each of five individuals will result in the equivalent of one QALY if the benefit is maintained over a 1-year period. This system has the advantage of considering both benefits and side-effects of programs in terms of the common QALY units. Although QALYs are typically assessed for patients, they can also be measured for others, including caregivers who are placed at risk because they experience excess stress and burden.

One of the important contrasts between the QWB and the SF-36 is that the QWB is primarily a preference-based measure designed to produce estimates of QALYs. The SF-36 is primarily a profile measure. However, summary scores are available for the SF-36 mental and physical health components. Similarly, profile information can be obtained from the QWB and QWB-SA. In fact, some investigators prefer to report profile information along with the overall score. For example, Patterson et al. [37] evaluated QWB outcomes for patients with late onset schizophrenia. The study is interesting because it is always assumed that the variation in outcomes for patients with mental illness will be reflected by mental health items. Figure 5 summarizes some profile information from the Patterson study. It compares patients who are normal, or have low, moderate, or high degrees of dysfunction associated with schizophrenia (evaluated by formal psychiatric interviews) in relation to QWB scales for...
Table 2 Principal concepts and domains of health-related quality of life contained in general preference weighted instruments for assessing quality-adjusted life years

<table>
<thead>
<tr>
<th>Concept</th>
<th>Disability</th>
<th>Distress</th>
<th>EuroQol</th>
<th>15D</th>
<th>Mark I</th>
<th>Mark II</th>
<th>Mark III</th>
<th>Years of Healthy Life</th>
<th>Quality of Well-being</th>
<th>Quality of Life and Health</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health perceptions</td>
<td>**</td>
<td>**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>**</td>
</tr>
<tr>
<td>Social function</td>
<td></td>
<td></td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social relations</td>
<td></td>
<td></td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Usual social role</td>
<td></td>
<td></td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intimacy/sexual function</td>
<td></td>
<td></td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
<td></td>
<td></td>
<td>**</td>
</tr>
<tr>
<td>Communication/speech</td>
<td></td>
<td></td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Psychological function</td>
<td></td>
<td></td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cognitive function</td>
<td></td>
<td></td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
<td></td>
<td></td>
<td>**</td>
</tr>
<tr>
<td>Emotional function</td>
<td></td>
<td></td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mood/feelings</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical function</td>
<td></td>
<td></td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mobility</td>
<td></td>
<td></td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
<td></td>
<td></td>
<td>**</td>
</tr>
<tr>
<td>Physical activity</td>
<td></td>
<td></td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
<td></td>
<td></td>
<td>**</td>
</tr>
<tr>
<td>Self-care</td>
<td></td>
<td></td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
<td></td>
<td></td>
<td>**</td>
</tr>
<tr>
<td>Impairment</td>
<td></td>
<td></td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
<td></td>
<td></td>
<td>**</td>
</tr>
<tr>
<td>Sensory function/loss</td>
<td></td>
<td></td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
<td></td>
<td></td>
<td>**</td>
</tr>
<tr>
<td>Symptoms/impairments</td>
<td></td>
<td></td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
<td></td>
<td></td>
<td>**</td>
</tr>
</tbody>
</table>


mobility, physical activity, and social activity. The scaling was the adjustment (from 1.0) in the QWB scale. Thus, higher scores in the figure represent more dysfunction. As the figure demonstrates, those with more severe schizophrenia have higher adjustments on the QWB subscales. The figure is interesting because it demonstrates that QWB profile information can be reported if the investigator desires. Further, the figure demonstrates that traditional functional scales capture variability associated with mental illness. We will review the issue of mental health in more detail in the last section of the paper.

**Comparisons between the QWB and SF-36: studies**

**Content and face validity**

In 1993, the US Department of Health and Human Services appointed a multi-disciplinary group of methodologists to recommend standardized strategies for the evaluation of health care. The panel, which released its report in 1996, suggested that standardized outcomes analyses be conducted to evaluate the cost-effectiveness of medical care [36]. These analyses require preference weighted measures of health related quality of life. Although there has been considerable interest in measuring the cost-effectiveness of treatments, little is known about the validity of general outcome measures and it is often difficult to choose between different approaches. Some authors have attempted to simplify the task by offering summary tables. An example of one such summary is shown as Table 2.

Despite the attractiveness of this approach, there are also some difficulties. In particular, creators of the tables typically examine the names of subscales, rather than the content of the measures. Consider the example of sensory function or loss. According to Table 2, sensory functioning is not included in the QWB. The creators of the table came to this conclusion because there is no subscale on the QWB named sensory function. However, the QWB-SA includes symptoms for loss of vision, loss of hearing, impairment of vision (including wearing glasses or contact lenses), problems with taste and smell, and so on. In fact the newer self-administered QWB (QWB-SA) devotes a major portion of the questionnaire to items on sensation and sensory organs. The symptoms include any hearing loss, blindness in one eye, blindness in both eyes, any problems with vision (floaters, double or distorted vision), eye pain, sensitivity to light, ear aches, difficulty in balance, and a variety of others. Indeed, the QWB-SA includes much more content on sensory functioning than do measures that are identified as including content on sensory functioning.
Correlations between the QWB and SF-36

A growing number of studies have applied both the SF-36 and the QWB. Table 3 summarizes correlations between the QWB and SF-36 components in four recent studies. The first study is based on a population sample in Beaver Dam, Wisconsin involving 1356 adults (560 men and 796 women). The mean age of the sample was 64.1 years [38]. In this sample, all of the participants lived in community settings. The second study involved 201 adults with serious illnesses, including 99 with AIDS, 74 with cancer, and 28 with other terminal illnesses [39]. The AIDS patients were younger (mean = 38 years) while those with cancer were older (mean age = 61 years). Those with other illnesses were also older (mean = 65 years). The third sample was a group of 100 HIV infected men (mean age 35.3 years). These men completed the MOS-HIV-34 which is a 34-item adaptation of the SF-36 for people with HIV disease [40]. The fourth sample included 301 community dwelling older adults (124 men and 177 women). All participants were older than 65 years and 44% of the men and 53% of the women were older than 75 years [41].

In all studies correlations between the QWB total score and the SF-36 physical function scale are substantial. Similarly, correlations between the QWB scale and the role-emotional scale are low in all four studies. As the table shows, correlations between the QWB and SF-36 components are quite consistent across these very different patient populations. The only exception is the low correlation between total QWB and bodily pain in the terminally ill sample. Overall, the available evidence suggests that the SF-36 and QWB tap most of the same variation in health status.

Ceiling effects

Measures differ in the extent to which they are responsive to minor variations in wellness. Some of our early studies demonstrated that nearly 80% of the general population has no functional limitations on a particular day [9]. Thus, functional items on surveys such as the National Health Interview Survey show the great majority of people to be well. On the other hand, only about 12% of the general population experience no symptom or problem during an average week [2]. Fryback et al. [38] studied the distributional properties of the QWB and SF-36 scales in Beaver Dam, Wisconsin. They found the QWB to be approximately normally distributed with scores of 1.0 being exceptionally rare. This finding has also been reported by our group [3] and others [41].

Ganiats et al. [42] have evaluated the QWB, SF-36, and other measures in clinical trials for patients with atrial fibrillation. Confirming other studies, they found 0% of patients at the ceiling level for the QWB. In contrast, many SF-36 scales were at their ceiling level. In other words, improvements on these dimensions would be difficult to detect because the highest values had already been recorded. For example, nearly 70% of respondents obtained the highest score for role-emotional and 54% obtained the highest score for social functioning. This was a concern because these patients were afflicted with a serious chronic illness. However, the SF-36 scores for this elderly population were comparable to those from the MOS normative sample (Figure 6).

One of the issues in the application of QWB, SF-36, and other measures is the percentage of respondents who have incomplete questionnaires. In the atrial fibrillation studies, all measures were administered to at least 300 patients. The percentage of patients who had any data missing for each measure was recorded. Because it was interviewer administered, no patient had incomplete survey responses for the QWB. The EQ-5D had 6% incomplete survey responses. The Health Utility Index (HUI) had 11% and the SF-36 had 26%. The new self-administered QWB also had some missing data problems similar to those of other measures. In the recent study by Andersen et al. [41], missing data for the QWB-SA were common: about 3% of the respondents failed to fill-out items on chronic diseases. Overall, the missing
data rate was about the same as it is for the SF-36. However, the SF-36 has worked out a scoring routine that allows the measures to be scored with up to 50% of the items missing. We are currently working on a missing data scoring protocol for the QWB-SA.

Mental health

Despite widespread interest in the concept of a QALY among practitioners in many different specialties, these ideas have received very little attention in the mental health fields. This reflects the widespread belief that mental health and physical health outcomes are conceptually distinct.

Although many questionnaires include different dimensions, they still may be tapping the same constructs. For example, a measure without a mental health component does not necessarily neglect mental health. Mental health symptoms may be included and the impact of mental health, cognitive functioning, or mental retardation may be represented in questions about role functioning. Some measures have multiple dimensions for mental health symptoms while others include fewer items that ask about problems in general. It is not clear that multiple measures are more capable of detecting clinical differences. This remains an empirical question for systematic analysis. A common strategy is to report outcomes along multiple dimensions.

Several years ago Kaplan and Anderson [43] argued that there are many similarities in mental health and physical health outcomes. The preference- and utility- based measures, that are assumed to ignore mental health, include the basic dimensions of observable functioning, symptoms, and duration. Mental health problems, like physical health problems, can be represented by symptoms and by disrupted role functioning. Consider some examples. Suppose that a patient has the primary symptom of a cough. If the cough does not disrupt role function, the preference or utility weighted score might show a small deviation from 1.0. If the cough is more serious and keeps the person at home, the score will be lower. If the cough is very severe, it might limit the person to a hospital and may have serious disruptive effects upon role functioning. This would necessitate an even lower score. Coughs can be of different duration. A cough associated with an acute respiratory infection may have a serious impact on functioning that may last only a short period of time. This would be indicated by a minor deviation in QALY's. A chronic cough associated with obstructive lung disease would be associated with significant loss of QALY's because duration is a major component of the calculation.

Now consider the case of a person with depression. Depression may be a symptom reported by a patient just as a cough is reported by other patients. Depression without disruption of role function would cause a minor variation of wellness. If the depression caused the person to stay at home the preference or utility weighted score would be lower. Severe depression might require the person to be in a hospital or special facility and would result in a lower score. Depressions, like coughs, are of different durations. Depression of long duration would cause the loss of more QALY's than would depression of short duration.

It is commonly asserted that the QWB excludes mental health content. Existing empirical evidence supports the validity of the QWB in studies of patients known to have impaired mental health. One study evaluated the validity of the QWB as an outcome measure for older psychotic patients. Seventy-two psychotic patients and 28 matched controls from the San Diego Veterans Affairs Medical Center completed the QWB; the Structured Clinical Interview for the DSM-III-R patient version, Scales for the Assessment of Positive and Negative Symptoms (SAPS and SANS), and the Global Severity Index (GSI) from the Brief Symptom Inventory were administered to all subjects. The QWB was significantly correlated with the SANS ($r = -0.52, P < 0.001$) the SAPS ($r = -0.57, P < 0.001$) and the GSI ($r = -0.62, P < 0.001$). There was a strong and significant linear relationship between QWB and severity of illness (as classified by the SANS and the SAPS). In addition, component scores of the QWB (i.e. mobility, physical activity, social activity, and worst symptom) were significantly lower among patients as compared to controls, and declined systematically as psychiatric symptoms increased [37]. In a related study, Rapaport et al. [44] found that patients with schizophrenia had QWB scores that were significantly lower (0.6) than age matched non-schizophrenia controls (0.7). QWB differences between these two groups were comparable in magnitude to a wide array of psychiatry specific measures.

Several other studies have now documented the sensitivity of the QWB to mental health problems. For example, a variety of studies has demonstrated the validity of the QWB for assessing depression among patients with HIV disease. In one study, ratings of depression using the Hamilton Depression Scale (HAM-D) were obtained from 285 HIV patients and 84 HIV negative men participating in the HNRC mixed gender cohort. The data were obtained at baseline and 6 months later. Depression was defined as Hamilton scores greater than 10. The study demonstrated a systematic relationship between Hamilton scores and QWB scores at baseline [45]. In addition, 22 HIV-positive subjects experienced increases of 10 points on the Hamilton Scale between the first evaluation and that at 6 months. For these individuals, significant reductions in QWB scores were observed. Analyses of QWB symptoms suggested greater symptom severity among those whose Hamilton scores increased. In addition, there was a greater reduction in physical activity. In other words, mood affected both symptoms and physical function.

Pyne et al. [20,21] compared QWB scores between patients with major depression and controls. The depressed patients were divided into mild, moderate and severe groups according to their scores on the HAM-D. There was a step-wise decrease in QWB scores for the controls through each of the patient groups. Similar results were observed for Beck Depression Inventory (BDI) scores. The difference between each of these groups is highly significant ($P < 0.001$). In order to evaluate which of six variables (age, sex, family history, presence of Axis III diagnosis, comorbid Axis I diagnosis, HAM-D) best predicted the QWB score in this patient population, a hierarchical multiple regression analysis was
conducted. Although Axis III disorders are significant predictors of QWB, depression (HAM-D and BDI) predicts QWB when Axis III diagnoses are controlled statistically.

The Beaver Dam study offers some evidence that the QWB is sensitive to mental health disorders in the community. The study offers QWB scores for individuals with a wide variety of conditions. Respondents with self-reported depression had the third lowest mean QWB score among all factor analyses, functioning items emerge as a separate factor.

The study offers QWB scores for individuals with a wide variety of conditions. Respondents with self-reported depression had the third lowest mean QWB score among all factor analyses, functioning items emerge as a separate factor.

The study offers QWB scores for individuals with a wide variety of conditions. Respondents with self-reported depression had the third lowest mean QWB score among all factor analyses, functioning items emerge as a separate factor.

QWBs can be used to compare treatments for physical and mental health problems. Suppose, for example, that a treatment for anxiety elevates patients from a level of 0.65 to a level of 0.75. Suppose further, that this treatment benefit lasted for 1 year. Each patient would gain 0.10 QALY (0.75–0.65 = 0.10 × 1 year = 0.10 QALY) for each year the benefit was observed. The treatment benefit would be expressed in terms of general QALY units. The productivity of the providers could be compared with providers in other areas of health care. All providers in health care use resources. Dividing the cost of a treatment by the QALY productivity provides the cost-utility ratio. Measuring productivity of mental health-related treatments in QALY units would allow the investments in mental health services to be compared directly to those in other aspects of health care.

One of the challenges to single-score measures is the argument that physical and mental health are different dimensions and that any measure combining them into a single index is like combining apples and oranges (CE Schwartz, RM Kaplan, JP Anderson, T Holbrook and MW Genderson, unpublished work). Clearly, mental and physical health services are different. Yet, providers compete for the same health care resources and there must be some basis for deciding how to allocate these resources.

We have conducted several studies designed to determine if mental and physical health clearly separate into well-defined dimensions. Ware et al. [31] have clearly shown that factor analyses of the SF-36 suggest separate physical and mental health dimensions. In a series of studies, we factor analyzed the QWB using groups of patients with multiple sclerosis (n = 263), non-insulin dependent diabetes mellitus (n = 420), non-head traumatic injury (n = 852), AIDS (n = 99), and cancer (n = 74). Each QWB item was treated as dichotomous and the matrix of 0s and 1s was subjected to factor analysis with orthogonal rotation. Items with factor loadings > 0.40 were regarded as consistent with a particular factor.

The analysis consistently revealed two strong factors. The first factor combined physical and psychological symptoms. The same factor emerged across these diverse patient populations. The items that consistently load on this factor include upset stomach, general tiredness, feeling upset, depressed or crying, dizziness, trouble sleeping, and excessive worry. In contrast to the symptom factor, the second dimension was defined by function. The second factor includes items such as spending the day in a wheelchair, difficulty in walking, and limitations in social activity (CE Schwartz, RM Kaplan, JP Anderson, T Holbrook and MW Genderson, unpublished work).

In three of these four populations, SF-36 data were also available. Using the same strategy, we have replicated factor analyses showing that mental and physical health separation for the SF-36. How might we explain why the QWB factor analysis separates physical health and mental health symptoms, whereas the SF-36 factor analysis separates them? One explanation is the difference between symptoms and functioning. In both factor analyses, functioning items emerge as a separate factor. The mental health items on the SF-36 use a different format and place greater emphasis on symptomatic activity. The SF-36 also separates instructions for role-emotional and physical functioning scales. Respondents are instructed to consider disruptions in activities that result from physical or mental health problems. Thus, separation of physical and emotional components in factor analysis might be expected because respondents had been prompted to think differently about them. It is worth noting that the separate dimensions of physical and mental health reported by Ware et al. [31] derive in part from their use of orthogonal rotations techniques. Methods that accommodate correlated factors (i.e. oblique rotation) have been shown to offer a better fit for health data [47].

Others have noted the complications resulting from the separation of physical and mental health. For example, Simon et al. evaluated 536 primary care patients before and after treatment with antidepressant medications. The mental health treatment was associated with improvements in both mental and physical health subscales of the SF-36 (physical function, role-physical, bodily pain, and general health perceptions). However, because of an artifact in the scoring system, the physical health summary score was unchanged. The study is important for two reasons: it demonstrates the complications in interpreting SF-36 summary scores; and it demonstrates that mental health treatments may have significant effects on measures believed to represent physical health [48]. We believe this underscores the fuzzy boundary between physical and mental health.

Summary

Outcomes researchers now have a variety of validated approaches to assess the costs, risks and benefits in medical care. The QWB and the SF-36 are two methods that arise from similar traditions. Many components of the measures are similar and correlations between the measures are substantial. Until recently, the SF-36 was much easier and less expensive to administer. However, an inexpensive self-administered form of the QWB is now available.

There are also important distinctions between the QWB and SF-36. The QWB places greater emphasis on symptoms and provides more clinical information. For example, it offers the clinician symptoms reports similar to a review of systems. Further, the QWB can be used for policy analysis because outcomes can be translated into QALYs. Although it is possible to create profiles from the QWB, it does less well than the SF-36 for characterizing multi-dimensional patterns of outcome. The QWB and SF-36 represent different
measurement approaches. The SF-36 is rooted in psychometric theory. The QWB, although adhering to some principles of psychometric theory, arises from a decision theory tradition. Because of the way it is constructed, some aspects of psychometric theory do not apply directly to the QWB. For example, traditional test–retest reliability has little value for assessing either the QWB or the SF-36. The QWB is better suited to policy analysis and to economic studies that require the calculation of a QALY. Investigators interested in cost-effectiveness or cost-utility analysis should consider using the QWB or a related utility-based measure. Investigators interested in reviewing a profile of outcomes may be better to use the well-established and well-validated SF-36.

We believe that better measurement technologies will replace many current approaches to health status assessment. Continuing research is necessary to build the next generation of outcome measures.

Acknowledgements

This work was supported in part by grants R01 HS 09170 and 5RO1NS33519 from the Agency for Health Care Policy and Research and grant P60 AR 40770 from the National Institutes of Health.

References


Accepted for publication 4 September 1998