Quality of Life Measurement After Stroke
Uses and Abuses of the SF-36

Jeremy C. Hobart, PhD; Linda S. Williams, MD; Kimberly Moran, MS; Alan J. Thompson, MD

Background and Purpose—The Medical Outcomes Study 36-item Short-Form Health Survey (SF-36) is widely used to measure health status after stroke. However, a fundamental assumption for its valid use after stroke has not been comprehensively tested: is it legitimate to generate scores for 8 scales and 2 summary measures using the standard algorithms? We tested this assumption.

Methods—SF-36 data from 177 people after stroke were examined (71% male; mean age, 62). We tested 6 scaling criteria to determine the legitimacy of generating the 8 SF-36 scale scores using Likert’s method of summed ratings, and we tested 2 scaling criteria to determine the appropriateness of the standard SF-36 algorithms for weighting and combining scale scores to generate 2 summary measures (physical and mental).

Results—Scaling assumptions were fully satisfied for 6 of the 8 scales, but 3 of these 6 scales had notable floor and/or ceiling effects. Assumptions for generating 2 SF-36 summary measures were not satisfied.

Conclusions—In this sample, 5 of the 8 SF-36 scales had limited validity as outcome measures after stroke, and the reporting of physical and mental summary scores was not supported. Results raise questions about the use of the SF-36 in stroke, and the SF-12 that is developed from it, and highlight the importance of testing scaling assumptions when applying existing scales to new populations. (Stroke. 2002;33:1348-1356.)

Key Words: health measurement ■ psychometrics ■ quality of life ■ SF-36 ■ stroke outcome

Recognition that healthcare evaluations should incorporate patients’ perspectives has led to the widespread use of patient-reported health rating scales as outcome measures. The purpose of these rating scales is to quantify, as rigorously as possible, aspects of health so that differences between people, changes over time, and changes associated with interventions can be detected accurately. Health rating scales are usually disease specific or generic. A key advantage of generic scales is that they enable comparisons across diseases. However, the use of generic rating scales assumes that they satisfy minimum psychometric requirements across diverse clinical populations.1,2 These assumptions are frequently untested. This is of concern because quality of inferences made from any study hinge directly on the scientific quality of the measurement instrument used.

The Medical Outcomes Study 36-item Short-Form Health Survey (SF-36) is a widely used, generic, patient-report, health status measure.3 It is recommended for use in health policy evaluations, general population surveys, clinical research, and clinical practice.4 In neurology, the SF-36 has been used in stroke (a MEDLINE search indicates more than 30 articles), motor neuron disease,5 Parkinson’s disease,6 epilepsy,7 headache,8 and multiple sclerosis.9–12 Moreover, it has been used frequently as a validating instrument in the psychometric evaluation of new measures.13–17

Two types of SF-36 scores can be generated (Figure 1). Scores for 8 scales are generated by summing items, without weighting or standardization, and scores for 2 summary measures are generated by combining weighted scale scores. The 8 scales provide a comprehensive profile of health status; however, the 2 summary measures have features that make them more advantageous for clinical trials. These features include better measurement precision, smaller confidence intervals, the elimination of floor and ceiling effects, simpler analysis by reducing the number of statistical tests required and avoiding the problem of multiple testing, and superior (theoretically) responsiveness.18,19 Summary measures are also more easily interpreted because their scores are directly related to scores for the general US population, which have been transformed to a mean of 50 and a standard deviation of 10.

Among the studies using the SF-36 in people with stroke, several have examined some of its psychometric properties. These studies report adequate internal consistency reliability20 and support the convergent and discriminant construct validity21 and group differences validity20 of the SF-36 in stroke. Floor and ceiling effects have been demonstrated by
some but not others, and the adequacy of completion rates in the elderly has been questioned. Despite these articles and comments that the SF-36 is a valid measure of health-related quality of life after stroke, no study has examined a fundamental prerequisite for rigorous measurement: tests of scaling assumptions. Tests of scaling assumptions determine whether it is legitimate to generate scale and summary scores using the algorithms proposed by the developers. Because psychometric properties are sample dependent, the performance of a measure in a specific application is more important than its performance generally. This study tested the scaling assumptions for the SF-36 in a sample of people with ischemic stroke.

**Subjects and Methods**

**Participants, Recruitment, and Data Collection**

People with ischemic stroke were recruited from 3 hospitals in Indianapolis between March 1999 and April 2001: an inner-city county hospital, a university-affiliated tertiary care hospital, and a Department of Veterans Affairs medical center. Patients were identified at the time of admission and, at follow-up, those who could communicate reliably (National Institutes of Health Stroke Scale Language score and no history of dementia) completed the SF-36 as part of a study to develop a disease-specific stroke scale (Stroke-Specific Quality of Life Scale). Demographic and clinical data were collected by interview and from medical records. Stroke severity on admission was determined using the retrospective Canadian Neurologic Scale. Low scores indicate greater impairment. SF-36 data were collected between 1 and 11 months after stroke in a face-to-face interview.

**The Measurement Model of the SF-36**

The measurement model of the SF-36 hypothesizes that 35 of the 36 items are grouped into 8 multi-item scales (physical functioning [PF], role limitations physical [RP], bodily pain [BP], general health perceptions [GH], energy/vitality [VT], social functioning [SF], role limitations emotional [RE], and mental health [MH]) that are aggregated into 2 summary measures (physical component [PCS] and mental component [MCS]). The remaining item is not used in scoring (Figure 1).

![Figure 1. Measurement model of the SF-36 (adapted from Ware et al19).](image-url)
difference by the US population standard deviation for that scale. Next, z scores are multiplied by their respective factor score coefficients, derived from US population data, and summed. Finally, these aggregated scores are standardized using a linear T score transformation to have a mean of 50 and standard deviation of 10, in the general US population.

The results of factor analytic studies led to the discovery of the 2 SF-36 summary measures. Principal components analysis (PCA) of intercorrelations among SF-36 scales from several studies consistently extracted 2 components, with similar scale-to-component correlations, that accounted for the majority of total SF-36 variance and variance in each of the individual SF-36 scales.18 These findings indicated that 2 summary measures could be generated without substantial loss of information. The 2 components were interpreted as measures of physical and mental health because the scales correlating highest with them were PF and RP, and MH and RE, respectively.

**Analyses**

**Is It Legitimate to Report SF-36 Scale Scores in Stroke?**

Six scaling assumptions must be satisfied for SF-36 scale scores to be generated using the item groups proposed by the developers and Likert's method of summed ratings.

1. Items in each scale must be roughly parallel (that is, measured at the same point on the scale and have similar variances) otherwise they do not contribute equally to the variance of the total score and must be standardized before combination.29 This criterion is evaluated by examining the equivalence of corrected item-total correlations. A range of minimum values has been recommended: 0.20, 0.30, and 0.40.1

2. Items in each scale must measure a common underlying construct otherwise it is not appropriate to combine them to generate a total score.26 This criterion is evaluated by examining the correlation between each item and the total score computed from the remaining items in that scale. This prevents overinflated values due to item overlap. These values are known as item-total correlations corrected for overlap, corrected item-total correlations, or item-total scale correlations. A range of minimum values has been recommended: 0.20, 0.30, and 0.40.1

3. Items in each scale should contain a similar proportion of information concerning the construct being measured otherwise they should be given different weights.26 This criterion is determined by examining the equivalence of corrected item-total correlations. Recently, Ware et al29 have stated that this criterion can be considered satisfied when values exceed 0.30, even if they vary.

4. Items must be correctly grouped into scales. That is, items must correlate substantially higher with the construct they are purported to measure than with the other constructs measured by the instrument.29 This criterion, termed item convergent and discriminant validity, is evaluated for each scale by comparing the difference between item-own and item-other scale correlations with 2 standard errors. Probable scaling failures are scored when item-own scale correlations exceed item-other scale correlations by more than 2 standard errors. Probable scaling failures are scored when item-own scale correlations exceed item-other scale correlations by less than 2 standard errors. Probable scaling failures are scored when item-other scale correlations exceed item-own scale correlations by less than 2 standard errors.

5. Scales must generate reliable estimates otherwise their scores cannot be confidently interpreted. This criterion is satisfied when Cronbach’s alpha coefficients30 for each scale exceed 0.7031 or 0.80.28 Cronbach’s alpha is a measure of reliability reflecting the (weighted) average correlation among all items in a scale (internal consistency).32,33

6. Scales must demonstrate that they measure distinct constructs otherwise interpretation of their scores is confounded. This criterion is satisfied when the correlations among scales are substantially less than their respective reliability estimates.29

**TABLE 1. Sample Characteristics (n=177)**

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age in years (mean (SD; range))</td>
<td>62 (13; 19–91)</td>
</tr>
<tr>
<td>Gender (male %)</td>
<td>71%</td>
</tr>
<tr>
<td>Race (White %)</td>
<td>73%</td>
</tr>
<tr>
<td>Stroke severity</td>
<td>9.5 (1.7; 2–11.5)</td>
</tr>
<tr>
<td>Discharge destination</td>
<td>72%</td>
</tr>
<tr>
<td>Timing of interview</td>
<td>3%</td>
</tr>
<tr>
<td>Rehabilitation status at time of interview</td>
<td>6%</td>
</tr>
<tr>
<td>Employment status at time of interview</td>
<td>9%</td>
</tr>
<tr>
<td>Not returned to work</td>
<td>33%</td>
</tr>
<tr>
<td>Not in paid employment at time of stroke</td>
<td>57%</td>
</tr>
<tr>
<td>SD indicates standard deviation; TOAST, Trial of Org 10172 in Acute Stroke Treatment.</td>
<td></td>
</tr>
</tbody>
</table>

**Is It Legitimate to Report SF-36 Summary Scores in Stroke?**

Two scaling assumptions must be satisfied for SF-36 PCS and MCS scores to be generated using the developers’ algorithm.5,18

1. PCA, with orthogonal (varimax) rotation of extracted components, of the correlations among the 8 SF-36 scales should support a 2D model of health that explains approximately 80% to 85% of the total reliable variance in the SF-36 scales and at least 75% of the reliable variance in each of the 8 SF-36 scales.3,18

2. The magnitude and pattern of correlations between the 8 SF-36 scales and the 2 rotated components should support their interpretation as measures of physical and mental health and be consistent with other studies. That is, the PCS measure should correlate strongly (>0.70) with the PF, RP, and BP scales and weakly (<0.30) with the MH and RE scales, and vice versa for the MCS measure.3,18

These 2 scaling assumptions were examined by undertaking a scale-level PCA with varimax rotation. Eigenvalues30 and the scree plot35 were examined to determine the optimum number of components to rotate.

**Results**

**Sample Characteristics**

A total of 177 people with stroke were studied. Table 1 reports the gender and race distribution, stroke subtype,
length of stay, stroke severity at onset, post-stroke rehabilitation and rehabilitation status at time of interview, and employment status at time of interview. There were more men than women, more whites than blacks, and almost half had small-vessel strokes (which have a better outcome). The stroke severity and rehabilitation data indicate that this is a sample of survivors from mild to moderate stroke.

Is It Legitimate to Generate SF-36 Scale Scores in Stroke?
All response options were endorsed for each item (Table 2), but endorsement frequencies were quite variable (range, 1.7% to 77.7%). Item response-option frequency distributions were relatively symmetrical for 6 of the 8 scales (PF, BP, GH, VT, SF, and MH), skewed toward worse health (low scores) for the RP scale, and skewed toward better health for the RE scale. Nevertheless, items within each of the 8 scales had similar mean scores and standard deviations, indicating they were roughly parallel. Skewness is explained below.

All item-own scale correlations corrected for overlap, except 2 items in the GH scale (11a and 11c), exceeded 0.40 indicating, for the other 7 scales, the items in each scale measured a common underlying construct, and that the criterion of Ware et al 29 for equivalence of item-total correlations was satisfied (Table 3). All item-own scale correlations exceeded item-other scale correlations, indicating no scaling failures. Most item-own scale correlations exceeded item-other scale correlations by a least 2 standard errors (>0.15), indicating definite scaling successes. However, in 2 scales, a substantial proportion of item-own scale correlations exceeded item-other scale correlations by less than 2 standard errors (<0.15), indicating high probable scaling success rates (SF = 43%, GH = 29%). This finding indicated that items in these 2 scales were limited in their ability to discriminate between constructs that are hypothesized to be different.

Alpha coefficients ranged from 0.68 to 0.90, indicating that most scales generated reliable scores (Table 4). One scale (GH) failed to satisfy the criterion of 0.70, and 2 scales (GH and SF) just failed to satisfy the more stringent criterion of 0.80. Intercorrelations among scales (range, 0.16 to 0.55) were substantially below their respective alphas, indicating that they were measuring 8 related but distinct constructs.

Scores for the 8 SF-36 scales spanned the entire, or almost entire (MH scale), scale ranges demonstrating good variability (Table 5). The shape of distributions were examined to determine the extent to which they were non-normal. Skewed distributions are those that are bunched up at one end. This is reflected in skewness statistics outside the recommended range of -1 to +1.36 Scores for the RP scale were positively skewed (skewness +1.18; distribution bunched to the left, low score end of the scale) indicating that respondents tended to be more physically disabled. In contrast, scores for the RE and MH scales were negatively skewed (skewness -0.84 and -0.67, respectively, distributions bunched to the right, high score end of the scale), indicating respondents had relatively better health in these domains. Scores for the other 6 scales were more evenly distributed (skewness -0.32 to +0.28). There were notable floor effects for 2 scales (RP 59.1%, RE 19.9%) and ceiling effects for 3 scales (RE 63.1%, SF 29.9%, BP 25.6%).

Is It Legitimate to Generate SF-36 Summary Scores in Stroke?
Principal components analysis of intercorrelations among SF-36 scales extracted 2 components with Eigenvalues exceeding unity, and the scree plot supported the existence of 2 higher order factors. This supported the hypothesis that a 2D model of health underpins the SF-36 in stroke. However, these 2 components explained less than 60% of the total reliable variance in all SF-36 scales and less than 75% of the reliable variable in 5 of the 8 scales (Table 6). Therefore, a substantial amount of information from SF-36 scales is lost when summary measures are reported in stroke. In addition, the magnitude and pattern of scale-to-component correlations in stroke differ from the US general population, indicating that the scale weights used to generate scores for the summary measures are not entirely applicable to people with stroke (Figure 2).

Discussion
This study has comprehensively examined the basic assumptions underpinning the scoring of the SF-36 in people with ischemic stroke. Scaling assumptions were not fully satisfied for either scale or summary scores. This has important clinical implications for the use of the SF-36, and the SF-12 that is derived from it.37,38 in stroke.

Results support the generation of summed scores for 6 of the 8 SF-36 scales (PF, RP, BP, VT, RE, and MH). However, GH and SF scales have low alphas and limited item convergent and discriminant validity, indicating imprecision and confounding of scores, respectively. In addition, 3 of the 6 scales for whom scaling assumptions were satisfied have important floor (RP, RE) and/or ceiling effects (BP, RE) that limit the value of these scales in clinical trials. The ceiling effect (percent sample at maximum score) represents a subsample of people whose score cannot increase regardless of any clinical improvement and for whom worsening of function may not register as a change in score. The floor effect (percent at minimum score) represents a subsample for whom clinical improvement may not register as a change in score and for whom worsening of function cannot be assessed. Therefore, floor and ceiling effects underestimate the impact of treatments in clinical studies, resulting in type II errors.

It is therefore important that the spectrum of health covered by the scale matches that of the study sample. This can be determined, to some extent, by examining the items and item response options of the scale, and the samples from whom and for whom the scale was developed. Even generic measures have some degree of specificity. For example, the SF-36 was developed to compare the effects of different healthcare financing arrangements and, consequently, is aimed toward less disabling medical conditions. Similarly, the Barthel Index, a generic measure of physical function, was developed for people with severe problems and is less appropriate for samples with milder physical disabilities.
TABLE 2. Descriptive Statistics for SF-36 Items (n=177)

<table>
<thead>
<tr>
<th>Scale</th>
<th>Item</th>
<th>Missing</th>
<th>Response Option Frequency Distribution (%)</th>
<th>Item Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>n</td>
<td>%</td>
<td>1</td>
</tr>
<tr>
<td>PF</td>
<td>3a</td>
<td>0</td>
<td>0</td>
<td>71.8</td>
</tr>
<tr>
<td></td>
<td>3b</td>
<td>0</td>
<td>0</td>
<td>38.4</td>
</tr>
<tr>
<td></td>
<td>3c</td>
<td>0</td>
<td>0</td>
<td>27.7</td>
</tr>
<tr>
<td></td>
<td>3d</td>
<td>1</td>
<td>0.6</td>
<td>45.5</td>
</tr>
<tr>
<td></td>
<td>3e</td>
<td>1</td>
<td>0.6</td>
<td>21.0</td>
</tr>
<tr>
<td></td>
<td>3f</td>
<td>1</td>
<td>0.6</td>
<td>27.3</td>
</tr>
<tr>
<td></td>
<td>3g</td>
<td>0</td>
<td>0</td>
<td>59.9</td>
</tr>
<tr>
<td></td>
<td>3h</td>
<td>0</td>
<td>0</td>
<td>38.4</td>
</tr>
<tr>
<td></td>
<td>3i</td>
<td>0</td>
<td>0</td>
<td>16.4</td>
</tr>
<tr>
<td></td>
<td>3j</td>
<td>0</td>
<td>0</td>
<td>7.3</td>
</tr>
<tr>
<td>RP</td>
<td>4a</td>
<td>2</td>
<td>1.2</td>
<td>71.4</td>
</tr>
<tr>
<td></td>
<td>4b</td>
<td>1</td>
<td>0.6</td>
<td>73.9</td>
</tr>
<tr>
<td></td>
<td>4c</td>
<td>1</td>
<td>0.6</td>
<td>76.7</td>
</tr>
<tr>
<td></td>
<td>4d</td>
<td>2</td>
<td>1.2</td>
<td>77.7</td>
</tr>
<tr>
<td>BP</td>
<td>7</td>
<td>1</td>
<td>0.6</td>
<td>26.7</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>1</td>
<td>0.6</td>
<td>48.3</td>
</tr>
<tr>
<td>GH</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2.8</td>
</tr>
<tr>
<td></td>
<td>11a</td>
<td>0</td>
<td>0</td>
<td>6.2</td>
</tr>
<tr>
<td></td>
<td>11b</td>
<td>0</td>
<td>0</td>
<td>11.3</td>
</tr>
<tr>
<td></td>
<td>11c</td>
<td>0</td>
<td>0</td>
<td>4.0</td>
</tr>
<tr>
<td></td>
<td>11d</td>
<td>0</td>
<td>0</td>
<td>5.1</td>
</tr>
<tr>
<td>VT</td>
<td>9a</td>
<td>0</td>
<td>0</td>
<td>6.2</td>
</tr>
<tr>
<td></td>
<td>9e</td>
<td>0</td>
<td>0</td>
<td>7.3</td>
</tr>
<tr>
<td></td>
<td>9g</td>
<td>0</td>
<td>0</td>
<td>6.2</td>
</tr>
<tr>
<td></td>
<td>9i</td>
<td>0</td>
<td>0</td>
<td>10.7</td>
</tr>
<tr>
<td>SF</td>
<td>6</td>
<td>1</td>
<td>0.6</td>
<td>40.9</td>
</tr>
<tr>
<td></td>
<td>9j</td>
<td>0</td>
<td>0</td>
<td>9.0</td>
</tr>
<tr>
<td>RE</td>
<td>5a</td>
<td>1</td>
<td>0.6</td>
<td>30.1</td>
</tr>
<tr>
<td></td>
<td>5b</td>
<td>1</td>
<td>0.6</td>
<td>32.4</td>
</tr>
<tr>
<td></td>
<td>5c</td>
<td>1</td>
<td>0.6</td>
<td>26.7</td>
</tr>
<tr>
<td>MH</td>
<td>9b</td>
<td>0</td>
<td>0</td>
<td>6.2</td>
</tr>
<tr>
<td></td>
<td>9c</td>
<td>0</td>
<td>0</td>
<td>2.8</td>
</tr>
<tr>
<td></td>
<td>9d</td>
<td>0</td>
<td>0</td>
<td>18.1</td>
</tr>
<tr>
<td></td>
<td>9f</td>
<td>0</td>
<td>0</td>
<td>1.7</td>
</tr>
<tr>
<td></td>
<td>9h</td>
<td>0</td>
<td>0</td>
<td>20.9</td>
</tr>
</tbody>
</table>

SF-36 indicates Medical Outcomes Study 36-item Short-Form Health Survey; SD, standard deviation; PF, physical functioning; RP, role limitations physical; BP, bodily pain; GH, general health perceptions; VT, energy/vitality; SF, social functioning; RE, role limitations emotional; MH, mental health.

*From raw questionnaires.
†Calculated after items scores reversed and transformed where appropriate.
‡This response option not available for this item.
Investigators should recognize that the term generic is relative and does not indicate universal applicability. Results do not support the computation of SF-36 PCS and MCS scores in stroke. Even though PCA supports a 2D model of health, this model does not explain as much of the variance in SF-36 scales as required, and the pattern of scale-to-component correlations is not consistent with findings in other clinical populations.\textsuperscript{2,39,40} Scale-to-component correlations for the GH and SF scales are the reverse of that required for the interpretation of the 2 components as measures of

\begin{table}
\begin{center}
\footnotesize
\begin{tabular}{|c|c|c|c|c|c|c|c|c|c|}
\hline
Scale & Item & PF & RP & BP & GH & VT & SF & RE & MH \\
\hline
PF & 3a & 0.59 & 0.49 & 0.28 & 0.26 & 0.45 & 0.42 & 0.17 & 0.13 \\
& 3b & 0.71 & 0.49 & 0.27 & 0.20 & 0.36 & 0.55 & 0.19 & 0.16 \\
& 3c & 0.71 & 0.50 & 0.31 & 0.22 & 0.35 & 0.52 & 0.28 & 0.19 \\
& 3d & 0.73 & 0.42 & 0.36 & 0.29 & 0.39 & 0.40 & 0.18 & 0.07 \\
& 3e & 0.67 & 0.31 & 0.34 & 0.28 & 0.27 & 0.30 & 0.15 & 0.11 \\
& 3f & 0.68 & 0.40 & 0.38 & 0.19 & 0.35 & 0.32 & 0.18 & 0.10 \\
& 3g & 0.69 & 0.39 & 0.33 & 0.22 & 0.31 & 0.35 & 0.19 & 0.06 \\
& 3h & 0.72 & 0.36 & 0.33 & 0.21 & 0.31 & 0.35 & 0.20 & 0.07 \\
& 3i & 0.57 & 0.31 & 0.30 & 0.23 & 0.23 & 0.31 & 0.21 & 0.12 \\
& 3j & 0.47 & 0.34 & 0.20 & 0.21 & 0.24 & 0.34 & 0.27 & 0.19 \\
RP & 4a & 0.50 & 0.69 & 0.33 & 0.13 & 0.36 & 0.44 & 0.29 & 0.22 \\
& 4b & 0.44 & 0.68 & 0.25 & 0.20 & 0.37 & 0.46 & 0.41 & 0.23 \\
& 4c & 0.45 & 0.71 & 0.31 & 0.21 & 0.34 & 0.40 & 0.31 & 0.24 \\
& 4d & 0.45 & 0.75 & 0.28 & 0.29 & 0.38 & 0.40 & 0.34 & 0.26 \\
BP & 7 & 0.37 & 0.27 & 0.74 & 0.23 & 0.33 & 0.25 & 0.16 & 0.25 \\
& 8 & 0.42 & 0.39 & 0.74 & 0.31 & 0.38 & 0.38 & 0.31 & 0.32 \\
GH & 1 & 0.32 & 0.16 & 0.18 & 0.53 & 0.28 & 0.22 & 0.29 & 0.23 \\
& 11a & 0.19 & 0.19 & 0.33 & 0.37 & 0.24 & 0.16 & 0.29 & 0.29 \\
& 11b & 0.24 & 0.23 & 0.13 & 0.45 & 0.32 & 0.09 & 0.21 & 0.13 \\
& 11c & 0.00 & 0.01 & 0.11 & 0.23 & 0.14 & 0.10 & 0.04 & 0.16 \\
& 11d & 0.29 & 0.22 & 0.23 & 0.61 & 0.47 & 0.31 & 0.28 & 0.33 \\
VT & 9a & 0.40 & 0.37 & 0.26 & 0.44 & 0.66 & 0.42 & 0.30 & 0.49 \\
& 9e & 0.43 & 0.43 & 0.35 & 0.43 & 0.75 & 0.45 & 0.29 & 0.49 \\
& 9g & 0.33 & 0.30 & 0.31 & 0.29 & 0.65 & 0.40 & 0.14 & 0.33 \\
& 9i & 0.31 & 0.33 & 0.35 & 0.30 & 0.71 & 0.44 & 0.19 & 0.36 \\
SF & 6 & 0.50 & 0.44 & 0.32 & 0.22 & 0.40 & 0.56 & 0.46 & 0.36 \\
& 10 & 0.44 & 0.46 & 0.29 & 0.25 & 0.49 & 0.56 & 0.30 & 0.33 \\
RE & 5a & 0.24 & 0.35 & 0.25 & 0.34 & 0.24 & 0.40 & 0.81 & 0.49 \\
& 5b & 0.23 & 0.40 & 0.22 & 0.33 & 0.25 & 0.41 & 0.81 & 0.45 \\
& 5c & 0.28 & 0.33 & 0.21 & 0.23 & 0.27 & 0.32 & 0.74 & 0.44 \\
MH & 9b & 0.06 & 0.18 & 0.24 & 0.21 & 0.37 & 0.22 & 0.24 & 0.49 \\
& 9c & 0.16 & 0.21 & 0.27 & 0.35 & 0.43 & 0.30 & 0.48 & 0.69 \\
& 9d & 0.05 & 0.23 & 0.12 & 0.22 & 0.32 & 0.25 & 0.35 & 0.57 \\
& 9f & 0.21 & 0.21 & 0.25 & 0.30 & 0.43 & 0.36 & 0.44 & 0.69 \\
& 9h & 0.14 & 0.22 & 0.28 & 0.22 & 0.36 & 0.36 & 0.41 & 0.54 \\
\hline
\end{tabular}
\end{center}
\caption{Item-to-Scale Correlations \((n=176-177)\)}
\end{table}

See Table 2 for abbreviations. Bolded values are item-own scale correlations (corrected for overlap). Nonbolded values are item-other scale correlations.
Figure 2 demonstrates that the scales do not cluster as in other studies. Table 4 indicates the correlations among the 4 scales hypothesized to correlate highest with the physical health component (PF, RP, BP, and GH; range, 0.25 to 0.55; mean, 0.36), and those hypothesized to correlate highest with the mental component (MH, RE, SF, VT; range, 0.28 to 0.51; mean, 0.43), are notably lower than those reported by others (PCS = 0.52 to 0.65, mean 0.57; MCS = 0.44 to 0.63, mean 0.54) and similar to those between physical and mental scales (range, 0.16 to 0.53; mean, 0.36). Consequently, relationships among the 8 health constructs measured by the SF-36 are disease specific.

This is a small study and its generalizability is uncertain. However, the SF-36 scale scores, their rank order, and the presence of floor and ceiling effects are similar to those reported in a large study of ischemic stroke patients for the UK collaborators of the International Stroke Trial.46 In addition, work from other neurological diseases demonstrates that results are consistent across diverse clinical sample.44 Nevertheless, the data indicate that this is a group of survivors of mild to moderate strokes. Severe strokes are underrepresented, as a result of language difficulty and difficulty returning to the clinics for appointments. Further studies are now essential to carefully evaluate the role of the SF-36 in stroke. This could be achieved from existing data sets.

It is also important to note that this study has only addressed scaling assumptions. These are fundamental crite-
ria that should be satisfied before more detailed psychometric evaluations are undertaken. More extensive evaluations of SF-36 scales are required to determine the extent to which they are valid and responsive indicators of the health constructs they purport to measure. Finally, this study demonstrates that stroke has direct or secondary effects on the major aspects of health (physical, psychological, and social function). Other health domains, not addressed by generic measures, are almost certainly affected by stroke. Although studies developing disease-specific rating scales are beginning to define these domains,24,47 the health impact of stroke and the patterns of change in health over time remain poorly understood.

The findings of this study provide the beginnings of evidence-based guidance for the use of the SF-36 in stroke. Until more extensive data are available, we recommend that SF-36 GH and SF scale scores should be interpreted with caution because there is evidence from this small study that neither are reliable and valid indicators. The RP, BP, SF, and RE scales might be poor choices as outcome measures for effectiveness and longitudinal studies because results from this study suggest they could underestimate health changes. Scores for the SF-36 PCS and MCS, and therefore the SF-12 from which only PCS and MCS scores can be constructed, cannot definitely be considered reliable and valid indicators of physical and mental health. This study also highlights the importance of comprehensive scale evaluation and underlines the need for caution when taking measures "off the shelf," expecting that psychometric attributes have been fully tested or that measures are generally applicable.

**Acknowledgments**

Dr Hobart is funded by a grant from the UK National Health Service Health Technology Assessment Programme, but the views and opinions expressed do not necessarily reflect those of the NHS Executive. Dr Williams is supported by a Research Career Development Award, Department of Veterans Affairs, Health Services Research and Development. This work was performed, in part, in the Regenstrief Institute for Health Care.

**References**

1. McHorney CA, Ware JE Jr, Lu JFR, Sherbourne CD. The MOS 36-Item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling

---

**TABLE 6. Correlations Between SF-36 Scales and Rotated Components**

<table>
<thead>
<tr>
<th>SF-36 Scale</th>
<th>Factor†</th>
<th>Stroke (n=177)</th>
<th>US Population (n=2474)*</th>
<th>Stroke (n=177)</th>
<th>US Population (n=2474)*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PCS</td>
<td>MCS</td>
<td>h²/r²t‡</td>
<td>PCS</td>
<td>MCS</td>
</tr>
<tr>
<td>PF</td>
<td>P</td>
<td>0.87</td>
<td>0.05</td>
<td>0.84</td>
<td>0.05</td>
</tr>
<tr>
<td>BP</td>
<td>P</td>
<td>0.75</td>
<td>0.21</td>
<td>0.71</td>
<td>0.21</td>
</tr>
<tr>
<td>GH</td>
<td>P&gt;M</td>
<td>0.59</td>
<td>0.25</td>
<td>0.48</td>
<td>0.76</td>
</tr>
<tr>
<td>VT</td>
<td>M&gt;P</td>
<td>0.54</td>
<td>0.60</td>
<td>0.62</td>
<td>0.69</td>
</tr>
<tr>
<td>SF</td>
<td>M</td>
<td>0.68</td>
<td>0.37</td>
<td>0.86</td>
<td>0.42</td>
</tr>
<tr>
<td>RE</td>
<td>M</td>
<td>0.23</td>
<td>0.71</td>
<td>0.63</td>
<td>0.17</td>
</tr>
<tr>
<td>MH</td>
<td>M</td>
<td>0.10</td>
<td>0.87</td>
<td>0.96</td>
<td>0.17</td>
</tr>
<tr>
<td>Eigenvalue</td>
<td></td>
<td>3.648</td>
<td>1.06</td>
<td></td>
<td>&gt;1.0</td>
</tr>
<tr>
<td>Variance</td>
<td></td>
<td>45.6</td>
<td>13.2</td>
<td></td>
<td>NR§</td>
</tr>
<tr>
<td>Total variance</td>
<td></td>
<td>58.8</td>
<td>81.5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*From Ware et al.
†Hypothesized factor content: P=physical factor content; M=mental factor content.
‡Total reliable variance in each SF-36 scale explained by the 2 principal components. h²=sum of squared factor loadings for each scale, r²t=alpha coefficient for each scale.
§Not reported.
¶Percent of total reliable variance in all SF-36 scales explained by the 2 principal components.

---

**Figure 2.** Plot of the SF-36 scale-to-component correlations. A, SF-36 in the US population (n=2474; prepared from Ware et al48). B, SF-36 in stroke (n=177).
Quality of Life Measurement After Stroke: Uses and Abuses of the SF-36
Jeremy C. Hobart, Linda S. Williams, Kimberly Moran and Alan J. Thompson

Stroke. 2002;33:1348-1356
doi: 10.1161/01.STR.0000015030.59594.B3
Stroke is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2002 American Heart Association, Inc. All rights reserved.
Print ISSN: 0039-2499. Online ISSN: 1524-4628

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://stroke.ahajournals.org/content/33/5/1348

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Stroke can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to Stroke is online at:
http://stroke.ahajournals.org//subscriptions/