

SF-36 Health Survey

Manual and Interpretation Guide

John E. Ware, Jr., Ph.D.

with

Kristin K. Snow, M.S.

Mark Kosinski, M.A.

Barbara Gandek, M.S.

The Health Institute, New England Medical Center • Boston, Massachusetts

© 1993

6. SCORING THE SF-36

This chapter provides scoring instructions for the eight multi-item scales and for the reported health transition item included in the SF-36 Health Survey. Chapter 3 describes the SF-36 scales and items. General scoring information and steps for data entry and scoring that are common to all items are discussed first (see Figure 6.1). Next, formulas for item aggregation and transformation of scale scores are presented. Finally, formal checks for errors in scoring are explained.

Importance of standardization

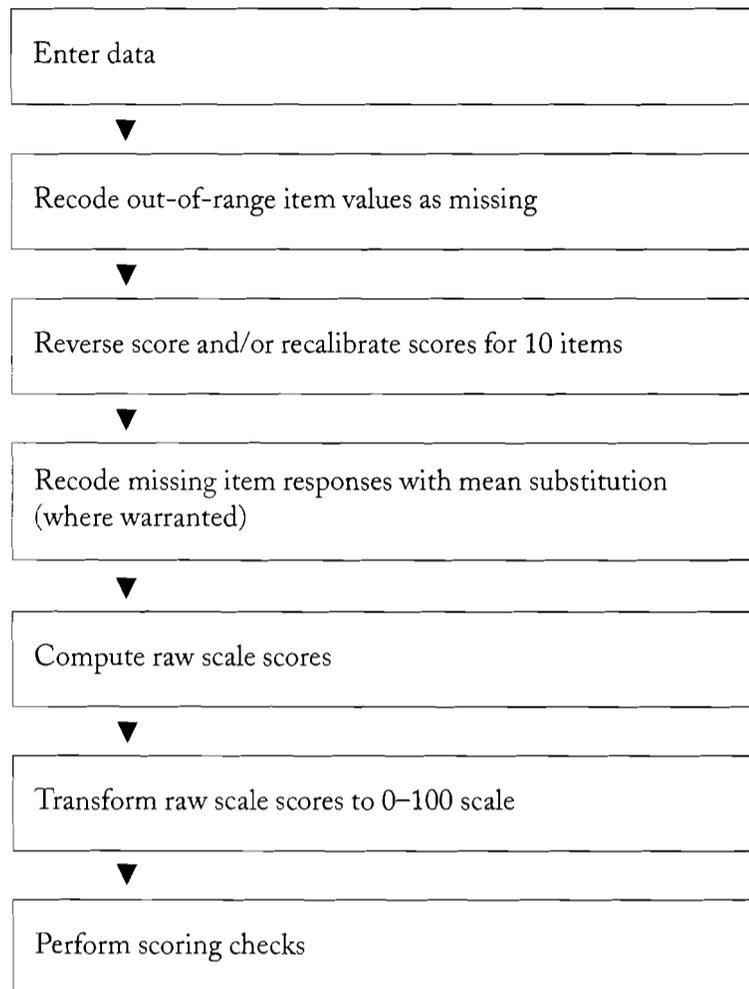
As with all standardized tests, standardization of content and scoring is what makes interpretation of the SF-36 scales possible. The content of the SF-36 form and the scoring algorithms were selected and standardized following careful study of many options. The algorithms described in this chapter were chosen to be as simple as possible while still satisfying the assumptions of the methods used to construct SF-36 scales.

Changes in the content of the survey or in scoring algorithms may compromise the reliability and validity of scores. Changes are also likely to bias scores sufficiently to invalidate normative comparisons and to prevent comparisons of results across studies.

There are at least two good reasons to adhere to the standards of content and scoring described in this manual. First, they are most likely to produce scores with the same reliability and validity as those reported here and in other Medical Outcomes Study (MOS) publications. Second, comparisons of results across studies are made possible to the benefit of all who use these content and scoring standards.

Prior to using the SF-36 scoring rules, it is essential to verify that the questionnaires being scored, including the questions asked (item stems), response choices, and numbers assigned to response choices at the time of data entry, have been reproduced exactly. The scoring rules described in this chapter are

FIGURE 6.1
FLOW CHART FOR
SCORING THE SF-36



appropriate for the standard SF-36 survey questions, response choices, and numbers assigned to response choices as reproduced in Appendix B. The chapter ends with algorithms that help to equate scores for the Developmental version and the Standard version of the SF-36.

General scoring information

SF-36 items and scales are scored so that a higher score indicates a better health state. For example, functioning scales are scored so that a high score indicates better functioning and the pain scale is scored so that a high score indicates freedom from pain. After data entry, items and scales are scored in three steps:

- (1) item recoding, for the 10 items that require recoding;
- (2) computing scale scores by summing across items in the same scale (raw scale scores); and
- (3) transforming raw scale scores to a 0 - 100 scale (transformed scale scores).

We recommend that both item recoding and scale scoring be performed by computer, using the scoring algorithms documented here or computer software available elsewhere (THI, 1992).

Data Entry

The SF-36 item responses should be keypunched as coded in the questionnaire. It is important to note that, although the numbers printed along with the response choices should be keypunched, they may not be the numbers ultimately assigned to those responses when SF-36 scales are scored.

In most cases, this means that the precoded number that is circled or marked by the respondent should be entered. However, sometimes it is not clear what number should be entered. Suggested rules for handling some of the more common coding problems are:

- If a respondent marks two responses which are adjacent to each other, randomly pick one and enter that number.

- If a respondent marks two responses for an item and they are not adjacent to each other, code that item “missing.”
- If a respondent marks three or more responses for an item, code that item “missing.”
- If a respondent answers the “yes/no” items by writing in “yes” or “no,” code the answer as though “yes” or “no” had been marked.

Response Technologies Inc. and other companies have developed scanning forms for use with the SF-36, in both standard and acute formats. Sample forms appear in Appendix B. Optical scanning generally reduces the time required to process questionnaires, but may involve greater initial investment in form design. Some scanning forms may require special processing equipment; however, this method may be cost-effective, especially if the SF-36 is being administered frequently or to a large sample (see Chapter 12).

Tables 6.1 through 6.9 present scoring information for the items used in each of the eight SF-36 health scales and the reported health transition item. Each table presents the verbatim content of each question, response choices, and both the precoded values printed in the questionnaire and final values for scoring each item. Item numbers in Tables 6.1 through 6.9 correspond to those on the Standard SF-36 form (reproduced in Appendix B).

Item Recoding

The next stage after data entry is the recoding of response choices as shown in Tables 6.1 through 6.9. Item recoding is the process of deriving the item values that will be used to calculate the scale scores. Several steps are included in this process: (1) change out-of-range values to missing, (2) recode values for 10 items, and (3) substitute person-specific estimates for missing items.

Out-of-Range Values

All 36 items should be checked for out-of-range values prior to assigning the final item values. Out-of-range values are those that are lower than an item’s precoded minimum value or higher than an item’s precoded maximum value (see Tables 6.1 through 6.9). Out-of-range values are usually caused by data-entry errors and, if possible, should be changed to the correct response through verification with the original questionnaire. If the questionnaire is not available, all out-of-range values should be recoded as missing data.

TABLE 6.11 PHYSICAL FUNCTIONING: VERBATIM ITEMS AND SCORING INFORMATION

Verbatim Items

- 3a. **Vigorous activities**, such as running, lifting heavy objects, participating in strenuous sports
- 3b. **Moderate activities**, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf
- 3c. Lifting or carrying groceries
- 3d. Climbing several flights of stairs
- 3e. Climbing one flight of stairs
- 3f. Bending, kneeling, or stooping
- 3g. Walking more than a mile
- 3h. Walking several blocks
- 3i. Walking one block
- 3j. Bathing or dressing yourself

Precoded and Final Values for Items 3a - 3j

<u>Response Choices</u>	<u>Precoded Item Value</u>	<u>Final Item Value</u>
Yes, limited a lot	1	1
Yes, limited a little	2	2
No, not limited at all	3	3

Scale Scoring

Compute the simple algebraic sum of the final item scores as shown in Table 6.11. See text for handling of missing item responses. This scale is scored so that a high score indicates better physical functioning.

Note. Precoded values are as shown on the appended form. This scale does not require recoding of items prior to computation of the scale score.

TABLE 6.2 ROLE-PHYSICAL: VERBATIM ITEMS AND SCORING INFORMATION

Verbatim Items

- 4a. Cut down the amount of time you spent on work or other activities
- 4b. Accomplished less than you would like
- 4c. Were limited in the kind of work or other activities
- 4d. Had difficulty performing the work or other activities (for example, it took extra effort)

Precoded and Final Values for Items 4a - 4d

<u>Response Choices</u>	<u>Precoded Item Value</u>	<u>Final Item Value</u>
Yes	1	1
No	2	2

Scale Scoring

Compute the simple algebraic sum of the final item values as shown in Table 6.11. See text for handling of missing item responses. This scale is scored so that a high score indicates better Role-Physical functioning.

Note. Precoded values are as shown on the appended form. This scale does not require recoding of items prior to computation of the scale score.

TABLE 6.11 BODILY PAIN: VERBATIM ITEMS AND SCORING INFORMATION

Verbatim Items

7. How much **bodily pain** have you had during the **past 4 weeks**?
8. During the **past 4 weeks**, how much did pain interfere with your normal work (including both work outside the home and housework)?

Precoded and Final Values for Item 7

Response Choices	Precoded Item Value	Final Item Value
None	1	6.0
Very mild	2	5.4
Mild	3	4.2
Moderate	4	3.1
Severe	5	2.2
Very severe	6	1.0

Scoring for Item 8— if both Items 7 and 8 are answered

Response Choices	If Item 8 Precoded Item Value	and Item 7 Precoded Item Value	then Item 8 Final Item Value
Not at all	1	1	6
Not at all	1	2 through 6	5
A little bit	2	1 through 6	4
Moderately	3	1 through 6	3
Quite a bit	4	1 through 6	2
Extremely	5	1 through 6	1

Scoring for Item 8— if Item 7 is not answered

Response Choices	Precoded Item Value	Final Item Value
Not at all	1	6.0
A little bit	2	4.75
Moderately	3	3.5
Quite a bit	4	2.25
Extremely	5	1.0

Scale Scoring

Compute the simple algebraic sum of final item values as shown in Table 6.11. See text for handling of missing item responses. This scale is scored positively so that a high score indicates lack of bodily pain.

Note. Precoded values are as shown on the appended form. This scale requires recoding of both items prior to computation of the scale score.

TABLE 6.4 GENERAL HEALTH: VERBATIM ITEMS AND SCORING INFORMATION

Verbatim Items

1. In general, would you say your health is:
 - 11a. I seem to get sick a little easier than other people
 - 11b. I am as healthy as anybody I know
 - 11c. I expect my health to get worse
 - 11d. My health is excellent

Precoded and Final Values for Items 1 & 11a-11d

Item 1	Response Choices	Precoded Item Value	Final Item Value
	Excellent	1	5.0
	Very good	2	4.4
	Good	3	3.4
	Fair	4	2.0
	Poor	5	1.0

Items 11a & 11c	Response Choices	Precoded Item Value	Final Item Value
	Definitely True	1	1
	Mostly True	2	2
	Don't Know	3	3
	Mostly False	4	4
	Definitely False	5	5

Items 11b & 11d	Response Choices	Precoded Item Value	Final Item Value
	Definitely True	1	5
	Mostly True	2	4
	Don't Know	3	3
	Mostly False	4	2
	Definitely False	5	1

Scale Scoring

Compute the simple algebraic sum of the final item values as shown in Table 6.11. See text for handling of missing item responses. This scale is scored so that a high score indicates better general health perceptions.

Note. Precoded values are as shown on the appended form. This scale requires recoding of three items prior to computation of the scale score.

TABLE 6.5 VITALITY: VERBATIM ITEMS AND SCORING INFORMATION

Verbatim Items

- 9a. Did you feel full of pep?
 9e. Did you have a lot of energy?
 9g. Did you feel worn out?
 9i. Did you feel tired?

Precoded and Final Values for Items 9a, 9e, 9g, & 9i

Items 9a & 9e	Response Choices	Precoded Item Value	Final Item Value
	All of the time	1	6
	Most of the time	2	5
	A good bit of the time	3	4
	Some of the time	4	3
	A little of the time	5	2
	None of the time	6	1

Items 9g & 9i	Response Choices	Precoded Item Value	Final Item Value
	All of the time	1	1
	Most of the time	2	2
	A good bit of the time	3	3
	Some of the time	4	4
	A little of the time	5	5
	None of the time	6	6

Scale Scoring

Compute the simple algebraic sum of the final item values as shown in Table 6.11. See text for handling of missing item responses. This scale is scored so that a high score indicates more vitality.

Note. Precoded values are as shown on the appended form. This scale requires recoding of two items prior to computation of the scale score.

TABLE 6.6 SOCIAL FUNCTIONING: VERBATIM ITEMS AND SCORING INFORMATION

Verbatim Items

6. During the past 4 weeks, to what extent has your physical health or emotional problems interfered with your normal social activities with family, friends, neighbors, or groups?
10. During the past 4 weeks, how much of the time has your physical health or emotional problems interfered with your social activities (like visiting with friends, relatives, etc.)?

Precoded and Final Values for Items 6 & 10

Item 6	Response Choices	Precoded Item Value	Final Item Value
	Not at all	1	5
	Slightly	2	4
	Moderately	3	3
	Quite a bit	4	2
	Extremely	5	1

Item 10	Response Choices	Precoded Item Value	Final Item Value
	All of the time	1	1
	Most of the time	2	2
	Some of the time	3	3
	A little of the time	4	4
	None of the time	5	5

Scale Scoring

Compute the simple algebraic sum of the final item values as shown in Table 6.11. See text for handling of missing item responses. This scale is scored so that a high score indicates better social functioning.

Note. Precoded values are as shown on the appended form. This scale requires recoding of one item prior to computation of the scale score.

TABLE 6.7 ROLE-EMOTIONAL: VERBATIM ITEMS AND SCORING INFORMATION

Verbatim Items

- 5a. Cut down the **amount of time** you spent on work or other activities
 5b. **Accomplished less** than you would like
 5c. Didn't do work or other activities as **carefully** as usual

Precoded and Final Values for Items 5a - 5c

<u>Response Choices</u>	<u>Precoded Item Value</u>	<u>Final Item Value</u>
Yes	1	1
No	2	2

Scale Scoring

Compute the simple algebraic sum of the final item values as shown in Table 6.11. See text for handling of missing item responses. This scale is scored so that a high score indicates better Role-Emotional functioning.

Note. Precoded values are as shown on the appended form. This scale does not require recoding of items prior to computation of the scale score.

TABLE 6.11
 PRECODED AND FINAL VALUES FOR ITEMS 9b, 9c, 9d, 9f, & 9h

Verbatim Items

- 9b. Have you been a very nervous person?
 9c. Have you felt so down in the dumps that nothing could cheer you up?
 9d. Have you felt calm and peaceful?
 9f. Have you felt downhearted and blue?
 9h. Have you been a happy person?

Precoded and Final Values for Items 9b, 9c, 9d, 9f, & 9h

Items 9b, 9c, & 9f	Response Choices	Precoded Item Value	Final Item Value
	All of the time	1	1
	Most of the time	2	2
	A good bit of the time	3	3
	Some of the time	4	4
	A little of the time	5	5
	None of the time	6	6

Items 9d & 9h	Response Choices	Precoded Item Value	Final Item Value
	All of the time	1	6
	Most of the time	2	5
	A good bit of the time	3	4
	Some of the time	4	3
	A little of the time	5	2
	None of the time	6	1

Scale Scoring

Compute the simple algebraic sum of the final item values as shown in Table 6.11. See the text for handling of missing item responses. This scale is scored so that a high score indicates better mental health.

Note. Precoded values are as shown on the appended form. This scale requires recoding of two items prior to computation of the scale score.

TABLE 6.9 REPORTED HEALTH TRANSITION: VERBATIM ITEM AND SCORING INFORMATION

Verbatim Item

2. Compared to one year ago, how would you rate your health in general now?

Precoded and Final Values for Item 2

Response Choices	Precoded Item Value
Much better now than one year ago	1
Somewhat better now than one year ago	2
About the same as one year ago	3
Somewhat worse now than one year ago	4
Much worse now than one year ago	5

Note. Precoded item values are as shown on the appended form. The average measured change in health for respondents selecting each response choice is presented in Chapter 9.

Recode Values for 10 Items

Seven items are reverse scored. Reverse scoring of items is done to ensure that a higher item value indicates better health on all SF-36 items and scales. SF-36 items that need to be reverse scored are worded so that a higher precoded item value indicates a poorer health state.

Item Recalibration

For 34 of the SF-36 items, research to date offers good support for the assumption of a linear relationship between item scores and the underlying health concept defined by their scales. However, empirical work has shown that two items require recalibration to satisfy this important scaling assumption. These items are in two different SF-36 scales: the General Health (GH) scale and the Bodily Pain (BP) scale.

General Health Rating Item. The “Very Good” and “Good” responses to Item 1 are recalibrated to achieve a better linear fit with the general health evaluation concept measured by the GH scale. Empirical studies during the Health Insurance Experiment (HIE) were among the first to document that the intervals between response choices for this item are not equal (Davies & Ware, 1981). Subsequent studies of Item 1, using both the Thurstone Method of Equal-Appearing Intervals (Thurstone & Chave, 1929) and other empirical methods, have also consistently shown that the interval between “Excellent” and “Very Good” is about half the size of the interval between “Fair” and “Good” (Ware, Nelson et al., 1992). These results have been confirmed in studies of SF-36 translations from 10 countries participating in the International Quality of Life Assessment (IQOLA) Project. Finally, in all studies we are aware of to date, mean values for a criterion general health scale for respondents who choose each of the five levels defined by Item 1 depart significantly from linearity.

Results from two MOS studies that served as the basis for the recommended recalibration of Item 1 are summarized in Table 6.10. As shown in Table 6.10 and discussed elsewhere (Ware, Nelson, et al., 1992), the mean criterion scores were remarkably similar for those who chose the same category of Item 1 across the screening (N=18,573) and longitudinal (N=3,054) samples. Intervals between adjacent response categories were unequal, as observed in the HIE (Davies & Ware, 1981). For these reasons, item scale values are transformed as shown in Table 6.10 using specific results from the

TABLE 6.10 MEAN CURRENT HEALTH SCORES FOR RESPONDENTS CHOOSING EACH LEVEL OF SF-36 ITEM 1

Response to Item 1	Mean Current Health		Recommended Scoring	
	Screening Sample (N=18,573)	Baseline Sample (N=3,054)	1-5 Scale	0-100 Scale
Excellent	87.9	86.9	5.0	100
Very good	75.5	75.4	4.4	84
Good	57.6	55.9	3.4	61
Fair	30.0	30.6	2.0	25
Poor	10.8	10.8	1.0	0

Note. Adapted from "Preliminary tests of a 6-item general health survey: A patient application" by J.E. Ware, E.C. Nelson et al., 1992, in A.L. Stewart & J.E. Ware (Eds.), *Measuring functioning and well-being: The Medical Outcomes Study approach* (p. 299). Durham, NC: Duke University Press.

screening sample. The result is a very high 0.70 correlation with the sum of the other four items in the GH scale.

Bodily Pain Items. The scoring rules recommended for the Bodily Pain (BP) scale were based on three considerations: (1) the items offer both different numbers and different content of response choices, (2) administration of Item 8 depended on the response to an item like Item 7 in the MOS, and (3) empirical studies indicate that recalibration of Item 7 is necessary to achieve a linear fit with the scale score and with other measures of bodily pain.

As shown in Table 6.3, the two bodily pain items offer an unequal number of response choices (six for Item 7 and five for Item 8). As a result, their variances are not equal, as required for a summated rating scale. Further, in all MOS studies published to date, Item 8 was administered (following a skip pattern) only to those respondents reporting at least some pain. Although the MOS skip pattern has been dropped to make the SF-36 easier to administer, the dependence between responses must be taken into account to compare results from new studies with published studies.

The recommended recoding of the first response choice for Item 8 on the basis of the response to Item 7 solves two problems. First, it converts Item

8 to a six-level item of roughly equal variance to Item 7. This is done by splitting those free of role interference due to pain into two different groups: (1) free of interference and free of pain (the best level), and (2) free of interference but with at least some pain (the next best level). Second, it approximates the dependence between the two items in MOS studies of reliability and validity to date (McHorney et al., 1992, 1993, in press).

Davies and Ware (1981) reported that recalibration of the bodily pain severity rating was necessary to satisfy the equal interval assumption in studies during the HIE. MOS studies have confirmed that the relationship between Item 7 and criterion measures of pain departs significantly from a linear association. Criterion pain measures used in these tests include visual analogue scales measuring pain severity and categorical ratings of pain frequency and duration. Final response values for Item 7 were derived from the mean values of a summary MOS criterion pain measure computed for respondents who chose each of the six levels defined by Item 7, using methods much like those illustrated in Table 6.10 for Item 1.

How to Treat Missing Data

Sometimes respondents leave one or more questionnaire items in a scale blank, although this happens infrequently (1 to 2% or less) in most surveys. One important advantage of multi-item scales is that a scale score can be estimated even though responses to some items are missing. Using a scoring algorithm that estimates missing values, it is usually possible to derive scale scores for nearly all respondents across the eight SF-36 scales.

We recommend that a scale score be calculated if a respondent answered at least half of the items in a multi-item scale (or half plus one in the case of scales with an odd number of items).

The recommended algorithm substitutes a person-specific estimate for any missing item when the respondent answered at least 50 percent of the items in a scale. A psychometrically sound estimate is the average score, across completed items in the same scale, for that respondent (Ware, Davies-Avery, & Brook, 1980). For example, if a respondent leaves one item in the 5-item Mental Health scale blank, substitute the respondent's average score (across the four completed mental health items) for that one item. When estimating the respondent's average score, use the respondent's final item values, as

defined in Tables 6.1 through 6.9. This step is easy to program using standard software packages (e.g., SPSS, SAS). Examples of program code and scoring software are available elsewhere (THI, 1992).

Computing Raw Scale Scores

After item recoding, including handling of missing data, a raw score is computed for each scale. This score is the simple algebraic sum of responses for all items in that scale, as shown in Table 6.11. For example, the raw scale score for the Role-Physical scale is the sum of the scores for Items 4a, 4b, 4c, and 4d. Use recoded items values and imputed values where applicable. Generally, we recommend that if the respondent answers at least 50% of the items in a multi-items scale, the score should be calculated. If the respondent did not answer at least 50% of the items, the score for that scale should be set to missing. Some prefer a more conservative approach for the scales with only two items and set those scales to missing unless both items are completed.

This simple scoring method is possible because items in the same scale have roughly equivalent relationships to the underlying health concept being measured, and no item is used in more than one scale. Thus, it is not necessary to standardize or weight items. These assumptions have been extensively tested and verified across 24 patient groups (McHorney et al., in press).

Transformation of Scale Scores

The next step involves transforming each raw scale score to a 0 to 100 scale using the formula shown below. Table 6.11 provides the information necessary to apply this formula to each scale.

$$\text{Transformed Scale} = \left[\frac{(\text{Actual raw score} - \text{lowest possible raw score})}{\text{Possible raw score range}} \right] \times 100$$

This transformation converts the lowest and highest possible scores to zero and 100, respectively. Scores between these values represent the percentage of the total possible score achieved. While this final step is optional, it is strongly recommended because transformed scale scores can be compared with norms derived from the MOS (McHorney et al., 1992, 1993, in press),

TABLE 6.II FORMULAS FOR SCORING AND TRANSFORMING SCALES

Scale	Sum Final Item Values (after recoding items as in Tables 6.1-6.8)	Lowest and highest possible raw scores	Possible raw score range
Physical Functioning	3a+3b+3c+3d+3e+ 3f+3g+3h+3i+3j	10, 30	20
Role-Physical	4a+4b+4c+4d	4, 8	4
Bodily Pain	7+8	2, 12	10
General Health	1+11a+11b+11c+11d	5, 25	20
Vitality	9a+9e+9g+9i	4, 24	20
Social Functioning	6+10	2, 10	8
Role-Emotional	5a+5b+5c	3, 6	3
Mental Health	9b+9c+9d+9f+9h	5, 30	25

Formula and example for transformation of raw scale scores

$$\text{Transformed Scale} = \left[\frac{(\text{Actual raw score} - \text{lowest possible raw score})}{\text{Possible raw score range}} \right] \times 100$$

Example: A Physical Functioning raw score of 21 would be transformed as follows:

$$\left[\frac{(21 - 10)}{20} \right] \times 100 = 55$$

Where lowest possible score = 10 and possible raw score range = 20

1990 National Survey of Functional Health Status, and other published and forthcoming results based on these scoring rules.

Raw and transformed scale scores are not calculated for the Reported Health Transition item. We recommend treating responses to this item as ordinal level data and analyzing the percentage of respondents who select each response choice or using the estimates of measured change reported for each response category in Chapter 9.

Scoring Checks

Because errors can occur while reproducing a form, entering data, programming or processing, which could lead to inaccurate scale scores, we strongly recommend formal scoring checks prior to using the scales. Any discrepancies observed during the following checks should be investigated for scoring errors:

- (1) Calculate SF-36 scale scores by hand for several respondents and compare the results to those produced by your scale-scoring computer software.
- (2) After items have been coded into their final item values, inspect the frequency distributions for the items to verify that only the final item values shown in Tables 6.1 through 6.9 are observed. Discrepancies should be limited to respondents with values estimated for missing data.
- (3) After items have been recoded and scale scores have been computed, inspect the correlation between each scale and its component items to verify that all correlations are positive in direction and substantial in magnitude (0.30 or higher).
- (4) Check correlations between the General Health scale and the other seven scales to verify that all are positive; with rare exceptions they should also be substantial in magnitude (0.30 or higher).
- (5) For those familiar with principal factor or components analysis, inspect correlations between the eight scales and the first unrotated factor or component extracted from the correlations among those scales. Regardless of extraction method, these correlations should be positive and substantial in magnitude (0.30 or higher).

Scoring of the SF-36 Developmental version

Some studies have been based on the Developmental version of the SF-36 made available in December 1988 (Ware, 1988). This section explains how to score the Developmental version to be more comparable with the Standard version. Thirty-five items across seven scales in the Developmental version can be scored identically to items in the Standard version.

One Social Functioning (SF) item differs in both item content and response choice format in the two versions (Item 9j in the Developmental version and Item 10 in the Standard version). The item content in the Standard version asks specifically if “physical health or emotional problems interfered with your social activities,” while the Developmental version asks if “health limited your social activities.” Further, only five response choices are provided in the Standard version to equate the variance of the two SF items without recalibration; thus, its scoring is simpler. The Developmental version had six response choices for this item.

To make scores for the SF scale in the Developmental version (Items 6 and 9j) more comparable to the Standard version: (1) reverse the scoring of the first item (Item 6); (2) recalibrate the second item (Item 9j) so it ranges from “1” to “5” rather than from “1” to “6”; and (3) compute the scale by summing the two items. Table 6.12 details these scoring steps.

Scoring alternatives

Scoring algorithms made available to users of the SF-36 Developmental version in 1988 are identical to those for the SF-36 Standard version for six of the eight scales (Ware, 1988). Both Developmental and Standard scoring algorithms include the recalibration of Item 1 of the GH scale, as documented and explained earlier in this chapter. We are not aware of any published studies in the United States, United Kingdom, or elsewhere that do not use the SF-36 scoring algorithm for the GH scale.

Those using the Developmental version of the SF-36 have a choice between Developmental (old) and Standard (new) scoring algorithms for the SF scale. The Developmental version of the second SF item (Item 9j) offered six

TABLE 6.12 SCORING THE SF-36 DEVELOPMENTAL VERSION SOCIAL FUNCTIONING SCALE

Verbatim Items

6. During the past 4 weeks, to what extent has your physical health or emotional problems interfered with your normal social activities with family, friends, neighbors, or groups?
- 9j. How much of the time during the past month has your health limited your social activities like visiting with friends or close relatives?

Precoded and Final Values for Items 6 & 9j

Item 6	Response Choices	Precoded Item Value	Final Item Value
	Not at all	1	5
	Slightly	2	4
	Moderately	3	3
	Quite a bit	4	2
	Extremely	5	1

Item 9j	Response Choices	Precoded Item Value	Final Item Value
	All of the time	1	1.0
	Most of the time	2	1.8
	A good bit of the time	3	2.6
	Some of the time	4	3.4
	A little of the time	5	4.2
	None of the time	6	5.0

Scale Scoring

Compute the simple algebraic sum of the final item values by summing Items 6 and 9j as described in the text. See text for handling of missing item responses. This scale is scored so that a high score indicates better social functioning.

Note. Precoded values are as shown on the appended form. This scale requires recoding of two items prior to computation of the scale score.

response choices, while the Standard version of this item (Item 10) has five response choices. The recalibration of this item, as recommended above (see “Scoring of the SF-36 Developmental version”), has no effect on the interpretation of SF scale scores. Instead, it adjusts the scale mean to be comparable to those in the normative data presented in Chapter 10.

Finally, we have published two options for scoring the Bodily Pain (BP) scale, which has identical item content across the Developmental and Standard versions. Advances in scoring for the Standard SF-36 are listed and explained above (see “Item Recalibration”). Early users of the Developmental version used the older (Developmental) scoring method (Ware, 1988). Thus, BP scales scored the old way will have means that are two to four points higher, on average, than BP scales scored the new (Standard) way. Scores for more than a third of respondents in the general population are shifted upward (i.e., towards better health) by five points or more when using the old scoring relative to the Standard scoring. The extent of this shift and implications for the precision and interpretation of BP scale scores, scored the old way, may vary depending on the proportion of respondents with chronic conditions and on their specific diagnoses. Therefore, we recommend routine use of the Standard scoring algorithms (as presented in this chapter). We encourage thorough documentation of any departures from this scoring system so that readers will know when they can and cannot compare results with other published studies.

Scoring Advances

We are presently evaluating several potential improvements in the scoring of the SF-36 including: (1) improvements in the enumeration of scale levels, (2) construction of aggregate (summary) indexes, and (3) norm-based scoring of scales and summary indexes. These and other SF-36 scoring issues that are likely to influence progress in the health assessment field are discussed in Chapter 12.