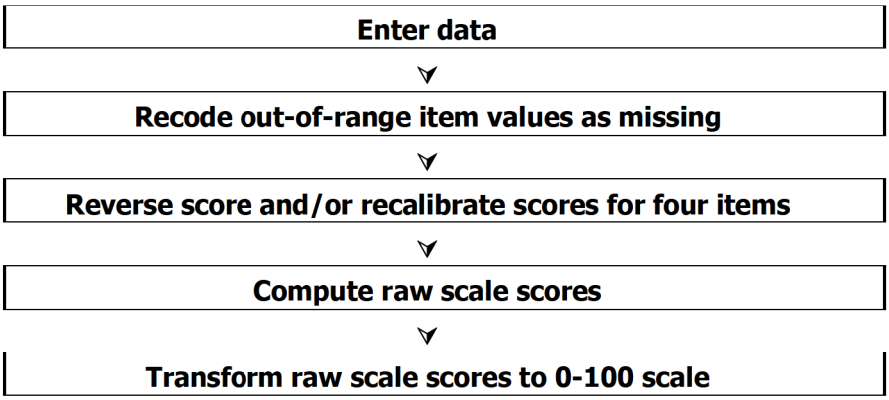
Chapter 6. How to Score SF-12 Health Survey Items

This chapter provides instructions for scoring SF-12v2 items and aggregating them into multi-item scales. General scoring information and steps for data entry and scoring that are common to all SF-12v2 items are discussed first (see Figure 6.1). Next, formulas for item aggregation and transformation of scale scores so that they will range from 0-100 are presented. These formulas apply to both standard (4-week recall) and acute (1-week recall) SF-12v2 survey forms. In the following chapters, an introduction to the NBS of SF-12v2 scales and summary measures is presented (Chapter 7), along with instructions for the NBS of scales and summary measures for the standard (Chapter 8) and acute (Chapter 9) survey forms.

Importance of Standardization

Standardization of content and scoring is what makes interpretation of SF-12v2 scales and summary measures possible. SF-12v2 item content and scoring algorithms were selected, developed and standardized following careful study of many options. The algorithms described in each of the scoring chapters were chosen to be as simple as possible while still satisfying the assumptions of the methods used to construct SF-12v2 scales and summary measures.

Figure 6.1 Flow Chart for Scoring SF-12v2 Items



General Scoring Information

SF-12v2 items are scored so that a higher score indicates a better health state. For example, functioning items are scored so that a high score indicates better functioning, and the pain item is scored so that a high score indicates freedom from pain. After data entry, items are scored in three steps:

(1) item recoding for the four 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).

Tables 6.2 through 6.9 present scoring information for the items used in each of the eight SF-12v2 health scales. Each table presents the 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 6.2 through 6.9 correspond to those on the standardized SF-12v2 standard and acute survey forms.

Item Recoding

The next stage after data entry is the recoding of response choices as shown in Tables 6.2 through 6.9. Item recoding is the process of deriving the item values that will be used to calculate the scale scores. Steps in this process include: (1) change out-of-range values to missing; and, (2) recode values for four items.

Out-of-Range Values

All 12 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 6.2 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.

Recode Values for Four Items

Four items are reverse scored. Reverse scoring of items is done to ensure that a higher item value indicates better health on all SF-12v2 items. SF-12v2 items that need to be reverse scored are worded so that a higher precoded item value indicates a poorer health state.

Item Recalibration

For 11 of the SF-12v2 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 one item (GH1, question 1) requires recalibration to satisfy this important scaling assumption.

General Health Rating Item

The "Very Good" and "Good" responses to Item 1 (GH1) 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 and Ware, 1981). Subsequent studies of Item 1 (GH1), using both the Thurstone Method of Equal-Appearing Intervals (Thurstone, 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, and Sherbourne, 1992). These results have been confirmed in studies of SF-36 translations from 10 countries participating in the IQOLA Project (Keller, Ware, Gandek et al., 1998). Finally, in all studies we are aware of to date, mean values for a criterion GH scale for respondents who choose each of the five levels defined by Item 1 (GH1) depart significantly from linearity.

Results from the MOS that served as the basis for the recommended recalibration of Item 1 (GH1) are summarized in Table 6.1. As shown in Table 6.1 and discussed elsewhere (Ware, Nelson, and Sherbourne, 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. For these reasons, item scale values are transformed as shown in Table 6.1 using specific results from the screening sample. The result is a very high correlation (0.70) with the sum of the other four items in the GH scale.

Table 6.1 Mean Current Health Scores for Respondents Choosing Each Level of SF-12v2 Item 1 (GH1)

	Mean Current Health		Recommended Scoring	
Response to Item 1	**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: *General Health Rating Item (GH1) is the first item in the General Health (GH) scale*

Table 6.2 Physical Functioning: Items and Scoring Information

Items

2a. Moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf

2b. Climbing several flights of stairs

Precoded and Final Values for Items 2a. and 2b.

Response Choices	*Precoded Item Value*	*Final Item Value*
Yes, limited a lot	1	1
Yes, limited a little	2	2
No, not limited at all	3	3

Raw Scale Scoring

Compute the simple algebraic sum of the final item scores as shown in Table 6.10. 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 6.3 Role Physical: Items and Scoring Information

Items

3a. Accomplished less than you would like

3b. Were limited in the kind of work or other activities

Precoded and Final Values for Items 3a. and 3b.

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

Raw Scale Scoring

Compute the simple algebraic sum of the final item scores as shown in Table 6.10. 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 6.4 Bodily Pain: Item and Scoring Information

Item

5. How much did pain interfere with your normal work (including both work outside the home and housework)

Precoded and Final Values for Item 5

Response Choices	*Precoded Item Value*	*Final Item Value*
Not at all	1	5
A little bit	2	4
Moderately	3	3
Quite a bit	4	2
Extremely	5	1

Raw Scale Scoring

Instructions for scoring the BP scale are presented in Table 6.10. See text for handling of missing item responses. This scale is scored 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 the item response values prior to computation of the scale score.*

Table 6.5 General Health: Item and Scoring Information

Item

1. In general, would you say your health is

Precoded and Final Values for 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

Raw Scale Scoring

Instructions for scoring the GH scale are presented in Table 6.10. 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 the item response values prior to computation of the scale score.*

Table 6.6 Vitality: Item and Scoring Information

Item

6b. Did you have a lot of energy

Precoded and Final Values for Item 6b.

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

Raw Scale Scoring

Instructions for scoring the vitality scale are presented in Table 6.10. 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 the item response values prior to computation of the scale score.*

Table 6.7 Social Functioning: Item and Scoring Information

Item

7. How much of the time has your physical health or emotional problems interfered with your social activities (like visiting friends, relatives, etc.)

Precoded and Final Values for Item 7

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

Raw Scale Scoring

Instructions for scoring the SF scale are presented in Table 6.10. 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 does not require recoding of the item response values prior to computation of the scale score.*

Table 6.8 Role Emotional: Items and Scoring Information

Items

4a. Accomplished less than you would like

4b. Did work or other activities less carefully than usual

Precoded and Final Values for Items 4a. and 4b.

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

Raw Scale Scoring

Compute the simple algebraic sum of the final item scores as shown in Table 6.10. 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 6.9 Mental Health: Items and Scoring Information

Items

6a. Have you felt calm and peaceful

6c. Have you felt downhearted and depressed

Precoded and Final Values for Items 6a. and 6c.

Item 6a. *Response Choices*	*Precoded Item Value*	*Final Item Value*
All of the time	1	5
Most of the time	2	4
Some of the time	3	3
A little of the time	4	2
None of the time	5	1
Item 6c. *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

Raw Scale Scoring

Compute the simple algebraic sum of the final item scores as shown in Table 6.10. See 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 one item prior to computation of the scale score.*

Computing Raw Scale Scores

After item recoding, 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 6.10. For example, the raw scale score for the RP Scale is the sum of the scores for Items 3a. and 3b. For single-item scales, the raw score is simply the final item response value. Use recoded item values where applicable.

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 (McHorney, Kosinski, and Ware, 1994).

Transformation of Scale Scores

The next step involves transforming each raw scale score to a 0-100 scale using the formula shown below. Table 6.10 provides the information necessary to apply this formula to each scale.

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

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.

Table 6.10 Scale Items Aggregated and Range of Possible Scores

SF-12v2 Scale	Sum Final Item Values (after recoding items as in Table 6.1-6.9)	Lowest and highest possible raw scores	Possible raw score range
Physical Functioning (PF)	Items 2a + 2b	2, 6	4
Role Physical (RP)	Items 3a + 3b	2, 10	8
Bodily Pain (BP)	Item 5	1, 5	4
General Health (GH)	Item 1	1, 5	4
Vitality (VT)	Item 6b	1, 5	4
Social Functioning (SF)	Item 7	1, 5	4
Role Emotional (RE)	Items 4a + 4b	2, 10	8
Mental Health (MH)	Items 6a + 6c	2, 10	8

Formula and example for transformation of raw scale scores to 0-100 scale scores:

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

Example: A Physical Functioning raw score of 5 would be transformed as follows:

$$\left[\frac{(5 - 2)}{4} \right] * 100 = 75$$

Where lowest possible score = 2 and possible raw score range = 4

Scoring Checks

Because errors can occur while reproducing a form, entering data, programming or processing, which could all 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 errors:

(1) calculate SF-12v2 0-100 scale scores by hand for several respondents and compare the results to those produced by your scoring program;

(2) after items have been coded into their final item values, inspect the frequency distributions of the items to verify that only the final item values shown in Tables 6.2 through 6.9 are observed; and,

(3) after items have been recoded and scale scores have been computed, inspect the correlation between each item and each scale to verify that all correlations are positive. With rare exceptions these correlations should also be substantial in magnitude (0.30 or higher).

How to Treat Missing Data

Although a simpler (sum score) method of MDE for *scale* scores derived from surveys with missing item responses appears to do as well in SF-12v2 studies to date, the recommended on-line scoring software for those scales and summary measures currently uses the same IRT-based method that we have adopted for the SF-36v2. There are several reasons for using the same MDE software that we use for the SF-36v2 with the SF-12v2 at this time.

First, regardless of which estimation method is used in scoring SF-12v2 scales with one or more missing responses, computerized software is required to estimate SF-12 PCS and MCS summary

measures when one of the eight scales is missing because of the number and complexity of the algorithms required. In such cases, the MDE on-line scoring software uses one or two of 14 regression models, seven models for each of the two summary scores.

Second, we believe that the IRT-based estimates of missing item responses will be required in the future for scale scores estimated from longer "static" forms, particularly for those with items that have a hierarchical relationship within a given scale (e.g., the SF-36® Health Survey PF scale). The "older" item averaging method may work as well as the IRT-based method for the SF-12 PF scale because both items in that scale define mid-range levels of PF. This hypothesis warrants formal testing.

Third, as the field advances, we expect to use the same IRT-based scale scoring algorithms for all "static" forms that estimate PF scores including those estimated from SF-8, SF-12, SF-36, and longer static PF forms. The same algorithms will be used to score responses to CAT–based dynamic administrations that individualize the selection of PF items from a calibrated item "pool". The same logic is being applied to long and short static forms and to dynamic forms that measure other generic and disease-specific domains of HR-QOL. In anticipation of the standardization of these metrics for widely-used measures of those domains, we have begun to standardize our item parameters and scoring algorithms at this time.

Fourth, to make results easier to interpret, we have adopted the NBS method for the SF-12v2 scales and summary measures and for all other estimates based on the generic item pools for those eight health domains. As discussed elsewhere (Kosinski, Bayliss, Bjørner et al., 2000; Ware, Kosinski, and Dewey, 2000; Ware and Kosinski, 2001), the advantages of NBS and standardization in relation to the SF-36 metrics are, respectively, that: (a) results from all scales and summary measures have the same direct interpretation in relation to their distributions of scores in the general U.S. population; and, (b) the improved score estimates have a direct relationship to, and therefore, can be interpreted in relation to guidelines derived from thousands of SF-36 studies already published in the peer-reviewed literature.

Finally, it should be noted that our on-line scoring software currently estimates both item and scale scores for the SF-12v2 for forms with missing data using the MDE algorithms discussed above. Information about this source is available on the Internet at http://www.qualitymetric.com.

Chapter 7. Introduction to Norm-Based Scoring

The interpretation of health status and outcomes has been made much easier with the NBS of SF-12 scales and summary measures. NBS makes it possible to compare and interpret the SF-12 profile and the physical and mental health summary measures. As originally documented in the first edition of this manual (Ware, Kosinski, and Keller, 1995), NBS is achieved by performing linear transformations of scores to achieve a mean of 50 and a *SD* of 10 in the general U.S. population, for both the SF-12 physical and mental health summary measures. This same transformation can be applied to all eight SF-12v2 scales to make comparisons across scales much more meaningful and to simplify their interpretation in relation to population norms.

We use 1998 norms to introduce NBS for the eight-dimension SF-12 profiles because these norms are much more up-to-date, and reflect differences in "physical health" observed in comparisons between 1998 and 1990 norms. These differences from 1990 to 1998 should be taken into account when drawing conclusions about population health, disease burden, and treatment benefits.

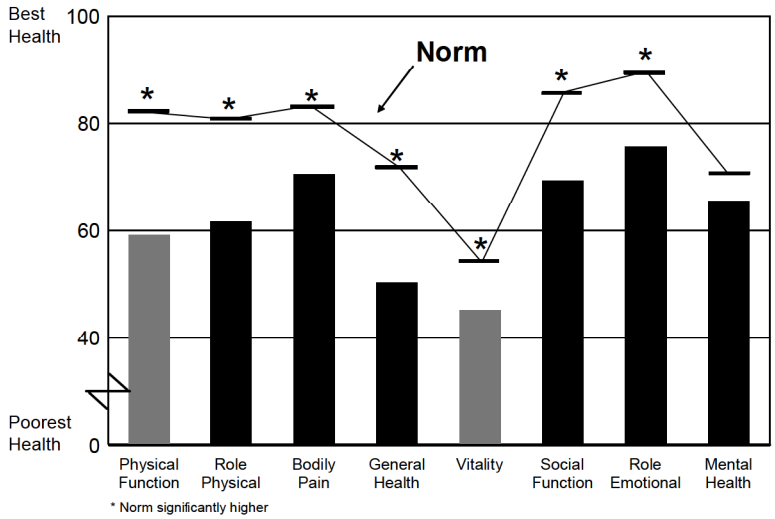
The use of 1998 norms and associated scoring algorithms for the SF-36, SF-12, and SF-8 also make it possible to cross-calibrate scores for the eight dimensions of health, which they all have in common, and for their physical and mental health summary measures. Finally, the 1998 norms and NBS algorithms provide the long awaited linkage necessary to compare results from administrations using SF-12 and those using SF-12v2.

Advantages of Norm-Based Scoring

The advantage of NBS can be illustrated by comparing a profile of SF-12 scales scored using 0–100 algorithms with the profile based on NBS algorithms for the same sample. For purposes of this comparison, we scored SF-12 profiles both ways for adults with diabetes (N=485) sampled from the general U.S. population in 1998. The 0-100 scoring of the eight SF-12 scales produced the profile shown in Figure 7.1. The shape of this profile – the peaks and valleys due to higher and lower scores across scales – reflect both the impact of diabetes, as well as arbitrary differences in the ceilings (scored 100) and floors (scored zero) of the SF-12 scales. For example, the VT scale measures a relatively wide score range and sets the ceiling relatively high by measuring very favorable levels of psychological well-being (Ware, Snow, Kosinski et al., 1993). Other scales, such as PF and RP, assess a narrower range based on a much lower ceiling defined as the absence of physical limitations. For these scales, scores of 100 have very different implications in terms of population norms. Ignoring these differences in norms and differences in the variances across scales, a reasonable inference from the profile in Figure 7.1 is that diabetes has a greater impact on VT (a score of about 45) than on PF (a score of nearly 60). This inference is incorrect. (See the two shaded scales.)

General population norms provide a basis for meaningful comparisons across scales (see Figure 7.1). For example, in the general population, the PF scale averages above 80 and the VT scale averages below 60 (on the 100-point score range). In relation to these norms, the impact of diabetes is actually much larger on the PF scale than on the VT scale, although both are statistically significant. Using the original 0–100 scoring, these differences in norms must be kept in mind when interpreting a profile. Differences in *SD*s, which are also substantial across some scales, must also be considered for purposes of comparing results across scales.

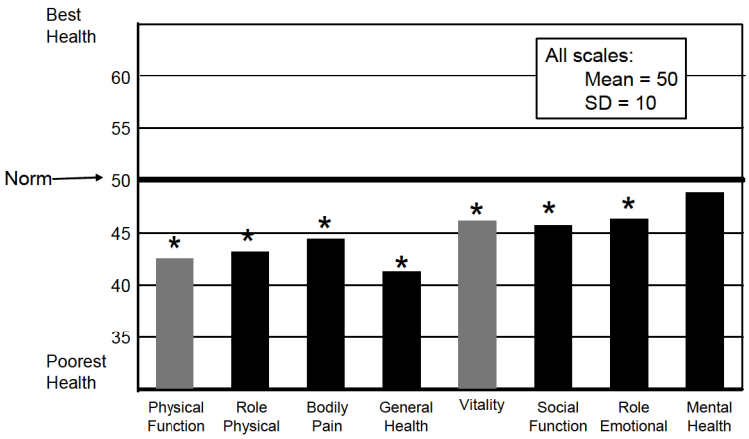
Figure 7.1 SF-12 Health Profile: Adults with Diabetes Compared with U.S. Norm



Using NBS, each scale is scored to have the same mean (50 points) and the same *SD* (10 points) in the general U.S. population in 1998. Without referring to tables of norms, it is clear with this method that anytime a scale score is below 50, health status is below average, and each point is one-tenth of a *SD*.

Figure 7.2 shows the same results for adults with diabetes using the simple linear NBS transformations. As shown in that figure, differences between the transformed scale scores and the population norm of 50, which is the same for all scales, much more clearly reflect the impact of the disease. Clinicians can more quickly see the decrements in seven of the eight scales in the SF-12 health profile that are significant for adults with diabetes.

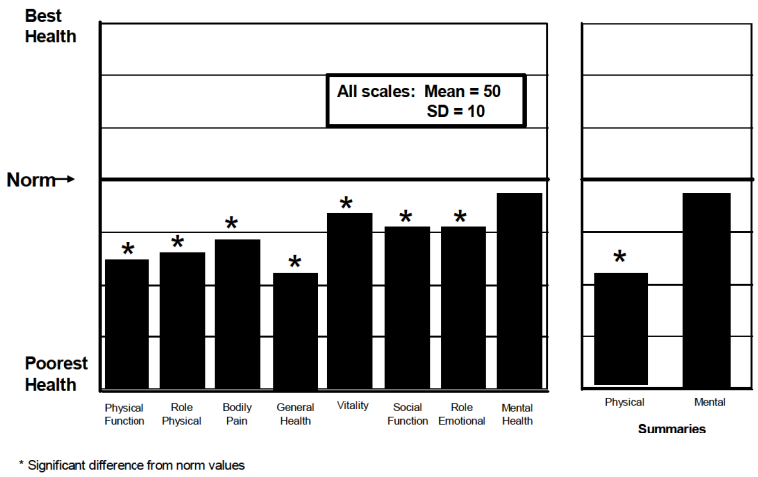
Figure 7.2 Norm-based Scoring of SF-12 Profile: Adults with Diabetes



Another noteworthy advantage of NBS is shown in Figure 7.3, namely the advantage of being able to meaningfully compare profiles and summary measures. Results for physical and mental summary measures, which have always been scored using NBS, can be compared directly with results for the eight SF-12 scales when all are standardized in relation to population norms.

Because the PCS and MCS measures take into account the correlations among the eight SF-12 scales, and only the PCS differs from the norm for adults with diabetes, it is clear from Figure 7.3 that diabetes has a very broad impact on the physical health profile and there is no independent effect on the mental summary score.

Figure 7.3 Interpreting SF-12 Profile and Summary Measures Among Adults with Diabetes



* Significant difference from norm values

Conclusion

In summary, the main advantage of NBS of the SF-12 is easier interpretation. When interpreting the scores you no longer have to remember the means and *SD*s for the eight scales and the two summary measures. In NBS, the general population norm is built into the scoring algorithm. All scores above or below 50 can be interpreted as above or below the general population norm. And, because the *SD*s for each scale are equalized at 10, it's easier to see exactly how far above (or below) the mean the score is in *SD* units. Furthermore, since all eight scales are scored to have the same *SD*, comparisons across scales and summary measures can be made directly.

NBS has another very important advantage. It provides a basis for direct comparisons between summary measure scores from the SF-12 and SF-12v2, because they were both normed and cross-calibrated in the 1998 general U.S. population. To facilitate such comparisons, Appendix C1 (Table C1.1) of this manual provides the information necessary to adjust the means for SF-12 PCS and MCS scores based on the original SF-12 algorithms to have a mean of 50 and a *SD* of 10 in the 1998 general U.S. population.

We conclude by underscoring the importance of carefully documenting the year and specific source of norms and scoring algorithms used in reports of "Study Methods" accompanying results when articles are based on the 1998 SF-12 U.S. population norms. Further, because tables and figures from journal articles are sometimes copied and distributed separately, we also recommend inclusion of explicit footnotes to "1998 SF-12 U.S. population norms" and to "NBS" in tables and figures presenting results based on the new 1998 U.S. population norms documented here.

Scoring the SF-12 Summary Measures – Translations

A question that often arises is whether to use standard (U.S.-derived) or country-specific SF-12 scoring algorithms in countries outside of the U.S. The appropriateness of using standard scoring algorithms for SF-12 has been demonstrated in nine countries. Large general population samples (n=1,413 to 6,124) from Denmark, France, Germany, Italy, the Netherlands, Norway, Spain, Sweden, and the United Kingdom were used to evaluate how well SF-12 summary measures, scored using both standard and country-specific scoring, replicated the SF-36 summary measures (Gandek, Ware, Aaronson et al., 1998). Product-moment correlations between comparable SF-36 and SF-12 summary measures (e.g., SF-36 PCS and SF-12 PCS) were very high, ranging from 0.94-0.96 and 0.94-0.97 for the physical and mental summary measures, respectively. This was true whether standard or country-specific scoring of the SF-36 or SF-12 measures was used. The mean values for comparable SF-36 summary measures and SF-12 summary measures were within 0.0 to 1.5 points (median=0.5 points) in each country, and were similar across age groups. Thus, within these countries, it is likely that standard and country-specific scoring of the SF-12 summary measures would lead to the same conclusions.

There are advantages and disadvantages to using either standard or country-specific scoring of the SF-12 summary measures. A major advantage of standard scoring is comparability: scores can be

compared across countries because they can be related to standard benchmarks, namely a mean of 50 and *SD* of 10 in the U.S. general population. Results also can be compared directly to the extensive normative data published in this manual and other documents. The disadvantage of standard scoring is that the mean health in all countries is not 50; for example, among adults age 18-74, mean MCS SF-12 scores were nearly three points higher in Denmark and Sweden than in the U.S. Thus, for studies within one country, comparisons are not being made to a norm of 50 for that country. There also are some countries (see Fukuhara, Ware, Kosinski et al., 1998) in which the standard scoring weights may not be appropriate due to differences from the U.S. in the underlying structure of the SF-36 and SF-12. For these countries, parallel analyses using standard and country-specific scoring may be appropriate, to determine if conclusions differ. If researchers do not use the standard scoring algorithms, they are strongly encouraged to precisely document what country-specific scoring was used, so readers may interpret their findings correctly. Additional information on country-specific scoring will be forthcoming on the IQOLA Web site at http://www.iqola.org and country-specific scoring modules will be made available on the QualityMetric Incorporated's Web site at http://www.qualitymetric.com.

Chapter 8. How To Score Standard Form Scales and Summary Measures

This chapter provides instructions for the NBS of the eight scales and two summary measures of the SF-12v2 standard (4-week recall) form (NBS was explained in Chapter 7). Instructions are provided below to linearly transform the 0-100 scale scores, as calculated from the instructions provided in Chapter 6, to have a mean of 50 and a *SD* of 10 in the 1998 general U.S. population. In addition, instructions are provided to aggregate the eight SF-12v2 scales into PCS and MCS measures. SF-12v2 summary measures and scales are scored using norm-based methods to have mean scores that are comparable to those for the SF-36v2.

Importance of Standardization

As with the scoring of SF-12v2 items, which are aggregated to score the scales, standardization of the scoring of the eight SF-12v2 scales and two summary measures is vital to their interpretation. Any changes in scoring of the SF-12v2 items, scales or the algorithms for the summary measures may compromise their reliability and validity. Changes in scoring have also been shown to invalidate normative comparisons, and changes are likely to complicate or prevent meaningful comparisons of results across studies.

Norm-Based Scoring of SF-12 Scale Scores (Standard 4-Week Recall)

The NBS of each SF-12v2 scale is accomplished using the formulas shown below. Table 8.1 provides the information necessary to apply these formulas to each SF-12v2 scale. The means and *SD*s (Table 8.1) used in NBS come from the 1998 general U.S. population. A linear z-score transformation is used so that all eight SF-12v2 scales have a mean of 50 and a *SD* of 10 in the 1998 general U.S. population.

The advantage of the standardization and NBS of the eight SF-12v2 scales is that results for one scale can be meaningfully compared with the other scales and their scores have a direct interpretation in relation to the distribution of scores in the 1998 general U.S. population. Specifically, all scores above or below 50 are above or below the mean, respectively, in the 1998 general U.S. population (see chapter 7 for more information). Because the *SD* is 10 for all eight scales, each one point difference or change in scores also has a direct interpretation. A one point difference or change is one-tenth of a *SD* unit or an effect size of 0.10. Lastly, NBS provides the basis for comparing the eight-scale scores with the two summary measure scores computed from the SF-12v2.

Step 1: Standardization of SF-12v2 Scales (Z-Scores), Standard Form

The first step in NBS consists of standardizing each SF-12v2 scale using a z-score transformation. A z-score for each scale is computed by subtracting the mean 0-100 score observed in the 1998 general U.S. population (Table 8.1) for each SF-12v2 scale score (0-100 scale) and dividing the difference by the corresponding scale *SD* (Table 8.1) from the 1998 general U.S. population. Formulas are listed below.

Formulas for z-score standardization of SF-12v2 scales, Standard Form:

PF_Z = (PF – 81.18122) / 29.10558
RP_Z = (RP – 80.52856) / 27.13526
BP_Z = (BP – 81.74015) / 24.53019
GH_Z = (GH – 72.19795) / 23.19041
VT_Z = (VT – 55.59090) / 24.84380
SF_Z = (SF – 83.73973) / 24.75775
RE_Z = (RE - 86.41051) / 22.35543
MH_Z = (MH – 70.18217) / 20.50597

Means and standard deviations are from Table 8.1

Table 8.1 1998 General U.S. Population Means and Standard Deviations Based on 0-100 Scoring Used to Derive SF-12v2 z-scores, Standard Form

SF-12 Scale	Mean	*SD*
Physical Functioning (PF)	81.18122	29.10558
Role Physical (RP)	80.52856	27.13526
Bodily Pain (BP)	81.74015	24.53019
General Health (GH)	72.19795	23.19041
Vitality (VT)	55.59090	24.84380
Social Functioning (SF)	83.73973	24.75775
Role Emotional (RE)	86.41051	22.35543
Mental Health (MH)	70.18217	20.50597

Step 2: Norm-Based Transformation of SF-12v2 Z-Scores, Standard Form

The second step involves linearly transforming each SF-12v2 z-score to the norm-based (50, 10) scoring. This is accomplished by multiplying each SF-12v2 z-score by 10 and adding the resulting product to 50. Formulas are listed below.

Norm-Based transformation of SF-12v2 z-scores, Standard Form:

Norm-Based PF: PF = 50 + (PF_Z * 10)
Norm-Based RP: RP = 50 + (RP_Z * 10)
Norm-Based BP: BP = 50 + (BP_Z * 10)
Norm-Based GH: GH = 50 + (GH_Z * 10)
Norm-Based VT: VT = 50 + (VT_Z * 10)
Norm-Based SF: SF = 50 + (SF_Z * 10)
Norm-Based RE: RE = 50 + (RE_Z * 10)
Norm-Based MH: MH = 50 + (MH_Z * 10)

Norm-Based Scoring SF-12v2 Physical and Mental Summary Measures (Standard 4-Week Recall)

Scoring of the SF-12v2 PCS and MCS measures involves three steps:

(1) First, the eight SF-12v2 scales are standardized using means and *SD*s for the 1998 general U.S. population.

(2) Second, they are aggregated using weights (factor score coefficients) from the 1990 general U.S. population. These are the same weights as those used to score PCS and MCS from the SF-36 (Ware, Kosinski, and Keller, 1994) and SF-36v2 (Ware, Kosinski, and Dewey, 2000).

(3) Finally, aggregate PCS and MCS scores are standardized using a linear t-score transformation to have a mean of 50 and a *SD* of 10, in the 1998 general U.S. population.

General U.S. population statistics used in the standardization and in the aggregation of SF-12v2 scale scores are presented in Table 8.1 and repeated in Table 8.2. Detailed information including formulas for scale aggregation and transformation of scores are also presented below. Formal checks using a test dataset made available upon request to QualityMetric Incorporated can be performed to confirm the successful reproduction of SF-12v2 PCS and MCS scales, as discussed later in the chapter. We strongly recommend these tests be used.

Norm-Based Scoring

The SF-12v2 PCS and MCS summary measures are scored using norm-based methods. The means and *SD*s used in scoring come from the 1998 general U.S. population and the factor score coefficients come from the 1990 general U.S. population (Ware, Kosinski, and Keller, 1994). A linear t-score transformation method is used so that both the PCS and MCS have a mean of 50 and a *SD* of 10 in the 1998 general U.S. population.

The advantage of the standardization and NBS of the PCS and MCS is that results for one summary score can be compared to the other and their scores have a direct interpretation in relation to the distribution of scores in the general U.S. population. Specifically, all scores above and below 50 are above and below the mean, respectively, in the 1998 general U.S. population. Because the *SD* is 10 for both PCS and MCS measures, each one point difference in scores also has a direct interpretation. A one point difference is one-tenth of a *SD*.

Table 8.2 1998 General U.S. Population Means, Standard Deviations and 1990 Factor Score Coefficients used to Derive PCS and MCS Scale Scores, Standard Form

SF-12v2 Scale	Mean*	*SD**	Factor Score Coefficients	
			PCS	MCS
PF	81.18122	29.10558	0.42402	-0.22999
RP	80.52856	27.13526	0.35119	-0.12329
BP	81.74015	24.53019	0.31754	-0.09731
GH	72.19795	23.19041	0.24954	-0.01571
VT	55.59090	24.84380	0.02877	0.23534
SF	83.73973	24.75775	-0.00753	0.26876
RE	86.41051	22.35543	-0.19206	0.43407
MH	70.18217	20.50597	-0.22069	0.48581

*__Note__: *The means and standard deviations for each SF-12v2 scale are based on the 0-100 scoring.*

Steps in Scoring

Following the scoring of the eight scales according to the standard SF-12v2 scoring algorithms (0-100 scale) explained in Chapter 6, the PCS and MCS are scored in three steps as explained below.

Step 1: Standardization of Scales (z-scores), Standard Form

First, each SF-12v2 scale is standardized using a z-score transformation using SF-12v2 scale means and *SD*s from the 1998 general U.S. population as given in Table 8.2. A z-score for each scale is computed by subtracting the 1998 general U.S. population mean from each SF-12v2 scale score (0-100 scale) and dividing the difference by the corresponding scale *SD* (0-100 scale) from the 1998 general U.S. population. Note that the SF-12v2 scales scored on the 0-100 scale are used in Step 1. Norm-based SF-12v2 scale

Formulas for z-score standardization of SF-12v2 scales, Standard Form:

PF_Z = (PF - 81.18122) / 29.10558
RP_Z = (RP - 80.52856) / 27.13526
BP_Z = (BP – 81.74015) / 24.53019
GH_Z = (GH – 72.19795) / 23.19041
VT_Z = (VT – 55.59090) / 24.84380
SF_Z = (SF - 83.73973) / 24.75775
RE_Z = (RE - 86.41051) / 22.35543
MH_Z = (MH – 70.18217) / 20.50597

Step 2: Aggregation of Scale Scores, Standard Form

After a z-score has been computed for each SF-12v2 scale, the second step involves computation of aggregate scores for the physical and mental summaries using the physical and mental factor score coefficients from the 1990 general U.S. population as given in Table 8.2.

Computation of an aggregate physical summary score consists of multiplying the z-score of each SF-12v2 scale by its respective physical factor score coefficient and summing the eight products, as shown below. Similarly, an aggregate mental summary score is obtained by multiplying the z-score of each SF-12v2 scale by its respective mental factor score coefficient and summing the eight products.

Formulas for aggregating scales in estimating aggregate physical and mental summary scores:

AGG_PHYS = (PF_Z * .42402) + (RP_Z * .35119) + (BP_Z * .31754) + (GH_Z * .24954) + (VT_Z * .02877) + (SF_Z * -.00753) + (RE_Z * -.19206) + (MH_Z * -.22069)

AGG_MENT = (PF_Z * -.22999) + (RP_Z * -.12329) + (BP_Z * -.09731) + (GH_Z * -.01571) + (VT_Z * .23534) + (SF_Z *.26876) + (RE_Z * .43407) + (MH_Z * .48581)

Step 3: Transformation of Summary Scores, Standard Form

The third step involves transforming the aggregate physical and mental summary scores to the norm-based (50, 10) scoring. This is accomplished by multiplying each aggregate summary score from Step 2 by 10 and adding the resulting product to 50. Formulas are listed below.

Formulas for t-score transformation of summary scores:
Transformed Physical (PCS) = 50 + (AGG_PHYS * 10)
Transformed Mental (MCS) = 50 + (AGG_MENT * 10)

Missing Data Estimation

The same psychometric models that make it possible to meaningfully estimate and compare scores for respondents who answer different sets of questions also make unbiased estimates of health scores possible even when some responses are missing. There are two theoretical advantages of using these models based on IRT to address the missing data problem. First, many scores that would have been missing using classical methods can be recovered. Second, in contrast to our original approach to estimating missing responses in the SF-12 and SF-36, an IRT-based approach should yield unbiased estimates of the PCS and MCS summary scores.

Results from ongoing evaluations of options for scoring PCS and MCS when a respondent is missing any one of the eight SF-12 or SF-36 scales have shown considerable promise in achieving both of the above objectives. For example, in re-analyses of data from the MOS the percentage of SF-36 scores that could be estimated for elderly participants was increased from 82.95% to 94.74% using IRT-based methods and our new scoring software with MDE algorithms (Kosinski, Bayliss, Bjørner et al., 2000). In both general and clinical populations, studies have shown that 50% of those who have missing PCS and MCS scores result from missing data on only one SF-12 scale. In the Medicare Health Outcomes Survey (NCQA, 2002), scores for nearly 40,000 respondents with one or more missing SF-36 responses at baseline in Cohort I were recovered and estimated without bias. For this reason, NCQA uses our new scoring software that incorporates MDE algorithms for its HEDIS Medicare outcomes survey (NCQA, 2002), and makes this scoring service available to the approximately 200 participating health care plans on the Internet.

More information about online scoring services with MDE algorithms for the SF-12 Health Survey is available on the Internet at http://www.qualitymetric.com.

(Alle metingen)
SF-12v2
Gezondheid
De volgende vragen gaan over uw standpunten t.a.v. uw gezondheid. Met behulp van deze gegevens kan worden bijgehouden hoe u zich voelt en hoe goed u in staat bent uw gebruikelijke bezigheden uit te voeren.

Hoe zou u over het algemeen uw gezondheid noemen? (General Health (GH))

- ❑ Uitstekend 5
- ❑ Zeer goed 4.4
- ❑ Goed 3.4
- ❑ Matig 2
- ❑ Slecht 1

GH vraag 1 (mogelijke score 1-5, mogelijke range is 4)

De volgende vragen gaan over bezigheden die u misschien doet op een doorsnee dag. Wordt u door uw gezondheid op dit moment beperkt bij deze bezigheden? Zo ja, in welke mate? (Physical Functioning (PF))

a) Matige inspanning, zoals het verplaatsen van een tafel, stofzuigen, zwemmen of fietsen

- ❑ Ja, ernstig beperkt 1
- ❑ Ja, een beetje beperkt 2
- ❑ Nee, helemaal niet beperkt 3

b) Een paar trappen oplopen

- ❑ Ja, ernstig beperkt 1
- ❑ Ja, een beetje beperkt 2
- ❑ Nee, helemaal niet beperkt 3

PF vraag 2a en 2b (mogelijke score 2-6, mogelijke range is 4)

Hoe vaak heeft u in de afgelopen 4 weken een van de volgende problemen bij uw werk of andere dagelijkse bezigheden gehad, ten gevolge van uw lichamelijke gezondheid? (Role Physical (RP))

a) U heeft minder bereikt dan u zou willen

- ❑ Altijd 1
- ❑ Meestal 2
- ❑ Soms 3
- ❑ Zelden 4
- ❑ Nooit 5

b) U was beperkt in het soort werk of andere bezigheden

- ❑ Altijd 1
- ❑ Meestal 2
- ❑ Soms 3
- ❑ Zelden 4
- ❑ Nooit 5

RP vraag 3a en 3b (mogelijke score 2-10 , mogelijke range is 8)

Hoe vaak heeft u in de afgelopen 4 weken een van de volgende problemen ondervonden bij uw werk of andere dagelijkse bezigheden ten gevolge van emotionele problemen (zoals depressieve of angstige gevoelens)? (Role Emotional (RE))

a) U heeft minder bereikt dan u zou willen

- ❑ Altijd 1
- ❑ Meestal 2
- ❑ Soms 3
- ❑ Zelden 4
- ❑ Nooit 5

b) U deed uw werk of andere bezigheden niet zo zorgvuldig als gewoonlijk

- ❑ Altijd 1

❑ Meestal 2
❑ Soms 3
❑ Zelden 4
❑ Nooit 5

RE vraag 4a en 4b (mogelijke score 2-10, mogelijke range is 8)

In welke mate bent u de afgelopen 4 weken door pijn gehinderd in uw normale werk (zowel werk buitenshuis als huishoudelijk werk)? (Bodily Pain (BP))

❑ Helemaal niet 5
❑ Klein beetje 4
❑ Nogal 3
❑ Veel 2
❑ Heel erg veel 1

BP vraag 5 (mogelijke score 1-5, mogelijke range is 4)

Deze vragen gaan over hoe u zich voelt en hoe het met u ging in de afgelopen 4 weken. Wilt u a.u.b. bij elke vraag het antwoord geven dat het best benadert hoe u zich voelde. Hoe vaak gedurende de afgelopen 4 weken…

a) Voelde u zich rustig en tevreden? (Mental Health (MH))

❑ Altijd 5
❑ Meestal 4
❑ Soms 3
❑ Zelden 2
❑ Nooit 1

c) Voelde u zich somber en neerslachtig? (Mental Health (MH))

❑ Altijd 1
❑ Meestal 2
❑ Soms 3
❑ Zelden 4
❑ Nooit 5

MH vraag 6a en 6c (mogelijke score 2-10, mogelijke range is 8)

b) Had u veel energie? (Vitality (VT))

❑ Altijd 5
❑ Meestal 4
❑ Soms 3
❑ Zelden 2
❑ Nooit 1

VT vraag 6b (mogelijke score 1-5, mogelijke range is 4)

Hoe vaak hebben uw lichamelijke gezondheid of emotionele problemen u gedurende de afgelopen 4 weken gehinderd bij uw sociale activiteiten (zoals vrienden of familie bezoeken, etc.)? (Social Functioning (SF))

❑ Altijd 1
❑ Meestal 2
❑ Soms 3
❑ Zelden 4
❑ Nooit 5

SF vraag 7 (mogelijke score 1-5, mogelijke range is 4)

Nu per schaal transformeren naar schaal scores (zijn er 8):

Getransformeerde schaal = ((echte score - laagst mogelijke schore)/range van score)) * 100

Pas op nu gaan we norm-based scores maken:
Per schaal een gemiddelde van 50 en een SD van 10 --> alle scores boven 50 liggen boven de norm van de algemene populatie en alle scores onder de 50 liggen onder de norm.

Stap 1 is de Z-score per schaal berekenen (hoeveel SD wijkt de schaalcore af van het gemiddelde): --> zie blz. 49 stap 1 --> doe dat per schaal. Het gemiddelde en de SD van schaal 8.1 zijn gebruikt

Stap 2 is samenvoegen van de de 8 schalen tot een mentale en fysieke schaal --> aggregatie met bepaalde weging (zie tabel 8.2) --> gebruik formule onderaan van pagina 49 stap 2.

Stap 3: nu moeten we nog van de mentale en fysieke schaal de t-score transformatie doen (zie formule op blz 50 bij stap 3.

Een score hoger of lager dan 50 betekent dan dat je hoger of lager scoort dan de algemene populatie. De score van 40 of 60 betekent dat je 1 SD lager of hoger scoort dan de algemene populatie.