

## Deciding Which Short Form Survey to Use

Choosing among the forms and versions of the SF family of health survey instruments depends on the requirements of the intended application, among other considerations. Score interpretation and the need for norms are no longer major considerations because the health domains and the underlying metrics (i.e., norm-based scoring) used in scoring all of the Short Form surveys have been standardized across the measures. In most cases, the choice involves a tradeoff between precision and respondent burden and whether computerized dynamic administrations are possible.

The sections that follow will focus on the SF instruments developed for use with adults.

### Features of the Short Form Surveys

**Original and revised versions.** Whereas the SF-36® Health Survey and SF-12® Health Survey are available in original and revised versions, the SF-8™ Health Survey is available in one version only. Although the SF-36v2™ Health Survey and SF-12v2™ Health Survey are very similar to their first version counterparts, each offers several improvements, including increased range and precision for the Role-Physical and Role-Emotional scales, improved item wording, and an easier-to-use format (these and other features are highlighted in the sections that follow). Because of these improvements, the SF-36v2™ Health Survey and SF-12v2™ Health Survey are recommended over the SF-36® Health Survey and SF-12® Health Survey, respectively, in most cases. Exceptions are noted below.

**Content.** All of the adult Short Form surveys measure the same eight health domains: Physical Functioning, Role-Physical, Bodily Pain, General Health,

Vitality, Social Functioning, Role-Emotional, and Mental Health. Because more items permit better representation of each health domain, the domains are best represented in the SF-36v2™ Health Survey and SF-36® Health Survey, followed by the SF-12v2™ Health Survey and SF-12® Health Survey, and then the SF-8™ Health Survey. (See Table 3.1 for a detailed summary of the descriptive and statistical characteristics of the SF-36v2™ Health Survey, SF-12v2™ Health Survey, and the SF-8™ Health Survey. A more detailed comparison of the SF-36® Health Survey with the SF-36v2™ Health Survey can be found in Chapter 13.) The improved question wording and simplified response categories of the SF-36v2™ Health Survey and SF-12v2™ Health Survey make these revised versions easier to understand and administer and less culturally biased than the original versions.

The SF-36® Health Survey and SF-12® Health Survey have 12 items in common, and this comparability was preserved in the updated versions of these two surveys. The SF-8™ Health Survey has only one item in common with the SF-36v2™ Health Survey and no items in common with the SF-12v2™ Health Survey. Content is very similar across all the surveys, however, and measures of corresponding concepts achieve a very high correlation across all forms. The SF-8™ Health Survey, SF-36® Health Survey, and SF-36v2™ Health Survey yield scores for the eight health domains as well as for the PCS and MCS measures. The SF-12v2™ Health Survey also produces the eight health domain scale and component summary measure scores, an improvement over the SF-12® Health Survey, which yields only component summary measure scores.

**Recall period.** Most of the items in each survey ask respondents to consider a specific period of time, or recall period, when responding. Both versions of the SF-36® Health Survey and SF-12® Health Survey are available in two forms, each covering a specific recall period. The *standard*, or 4-

week recall, form asks the respondent to answer the Short Form questions as they pertain to the way he or she felt or acted *during the past 4 weeks*. The *acute*, or 1-week recall, form asks the respondent to answer the Short Form questions as they pertain to the way he or she felt or acted *during the past week*. The SF-8™ Health Survey is available in three forms, each of which has been validated, with differing recall periods: the standard form uses a 4-week recall, an acute form that uses a 1-week recall, and a second acute form that uses a 24-hour recall (Ware, Kosinski, Dewey, & Gandek, 2001).

Use of the standard, or 4-week recall, form of the SF-36v2™ Health Survey is appropriate for cases in which the instrument will be administered only once to the respondent, or when at least 4 weeks will pass between a re-administration of the instrument. In most cases, the standard version will meet the needs of the clinician for patient monitoring and the needs of the researcher for many types of investigations, particularly those of a longitudinal nature.

The acute, or 1-week recall, form provides a better description of health status during the most recent week than the standard form. When more frequent re-administration is required, the acute form is most appropriate. For example, the acute form is recommended when a clinician or researcher wants to closely monitor the effects of a physical (e.g., pharmacological) or behavioral (e.g., psychotherapeutic) intervention on a patient or group of patients when such effects are likely to occur rapidly (e.g., asthma therapy). However, one or more weeks must pass between administrations of the acute version in order to obtain valid information.

Generally, the results from administrations of the standard and acute forms substantially agree. However, users may find that results from the acute form will differ from those obtained from the standard form. Keller et al. (1997), for example, found the effect of the form did

approach significance ( $p = .08$ ) with two small samples of asthma patients participating in a controlled study of the effects of inhaled corticosteroid on HRQOL. In addition, univariate analyses revealed more favorable results (higher scores on the 0–100 scoring metric) from the acute form, with RE averaging nearly 7 points higher ( $p = .05$ ), RP averaging nearly 5 points higher, and SF nearly 3 points higher. It is important, however, to note that this study was conducted within the context of a randomized clinical trial where changes in health status can occur relatively quickly, and, thus, it needs to be replicated with other acutely ill patient samples. Also, the Keller et al. findings could not be replicated with data from the SF-36v2™ Health Survey 1998 normative sample, where health domain scale scores from the standard and acute forms were very similar.

**Respondent burden.** Shorter surveys can be completed more quickly and require less space in printed questionnaires. The SF-8™ Health Survey can be completed in 1 to 2 minutes, on average. The SF-12® Health Survey and SF-12v2™ Health Survey require 2 to 3 minutes, on average, and the two versions of the SF-36® Health Survey require between 5 and 10 minutes, on average. Survey length and respondent burden may be an issue in some clinical settings or when the survey is administered as part of a large battery of instruments. Consequently, the SF-12v2™ Health Survey quickly became the tool of choice among fixed-form population surveys because its RP and RE health domain scales cover wider ranges of health levels more accurately with *fewer* items than their three- and four-item counterparts on the SF-36® Health Survey. This improvement in precision, in conjunction with a reduction in respondent burden, is noteworthy in light of the importance of the role participation domains and the increasing importance of practical considerations in selecting health measures for widespread use.

**Precision.** Like respondent burden,

precision in part varies directly with the numbers of items and response choices. The SF-8™ Health Survey scales are the coarsest, offering the least amount of precision and generally covering a narrower range of each of the eight health domains, and the longer SF-36v2™ Health Survey offers a greater degree of precision than the SF-12v2™ Health Survey. Furthermore, scales with more levels provide greater measurement precision (see Table 3.1). Across all domains, the SF-36v2™ Health Survey health domain scales have as many or more levels, and thus greater measurement precision, than any of the SF-12v2™ Health Survey or SF-8™ Health Survey scales. This is an important feature to consider when sample sizes are small and measurement precision is paramount. The improvements in the SF-36v2™ Health Survey and SF-12v2™ Health Survey significantly increased the precision of both of these surveys; as a result, the difference between these updated surveys is smaller than the difference between the SF-36® Health Survey and SF-12® Health Survey.

Note that the component summary measures of each of the Short Form instruments provide the greatest number of levels of measurement and, thus, more measurement precision than each of their form's health domain scales. For this reason, even the SF-8™ Health Survey component summary measures may provide sufficient measurement precision for studies involving small sample sizes.

**Treatment of missing data.** Two procedures have been developed for estimating scores when there are missing data within the Short Form surveys: the *Half-Scale Rule* and the *Full Missing Data Estimation (Full MDE)* (see Chapter 6). These procedures can be applied to data from all of the Short Form surveys; however, the most robust treatment of missing data occurs with the SF-36® Health Survey and SF-36v2™ Health Survey, followed by the SF-12® Health Survey and

SF-12v2™ Health Survey, and, finally, the SF-8™ Health Survey. Note that the Full MDE method requires the use of the QualityMetric Health Outcomes™ Scoring Software 2.0 (Saris-Baglana et al., 2007; see Chapter 5).

**Data quality evaluation (DQE).** Several measures and procedures have been developed or are otherwise available for evaluating the quality of data obtained from the administration of the Short Form surveys, including completeness of data, responses outside of range, confirmation of the two-component structure, percentage of estimable component scores, convergent validity, discriminant validity, consistent responses, percentage of estimable scale scores, item internal consistency, item discriminant validity, and scale reliability. Only some of these measures and procedures can be used with individual Short Form instruments (see Table 3.2; see also Chapter 6).

**Ceiling and floor effects.** Another major consideration when choosing among the Short Form surveys is ceiling and floor effects. With the exception of the RP and RE scales, the range of observed scores is greatest among SF-36® Health Survey and SF-36v2™ Health Survey health domain scales compared to SF-12® Health Survey, SF-12v2™ Health Survey, and SF-8™ Health Survey scales, although the differences are not great (see Table 3.1). The implication is that the SF-36® Health Survey and SF-36v2™ Health Survey health domain scales define a wider range of each construct measured than the SF-12® Health Survey, SF-12v2™ Health Survey, and SF-8™ Health Survey scales. Therefore, the ceiling and floor effects found with SF-36® Health Survey and SF-36v2™ Health Survey scales are less problematic. Moreover, with the incorporation of the revised role functioning items, the SF-36v2™ Health Survey is even less susceptible to these effects than the SF-36® Health Survey.

**Norms.** More comprehensive norms are

now available for the standard and acute forms of the SF-36v2™ Health Survey, SF-36® Health Survey, SF-12v2™ Health Survey, SF-12® Health Survey, and SF-8™ Health Survey. Norms for both versions of the SF-36® Health Survey and the SF-12® Health Survey are based on a 1998 U.S. general population sample, while the SF-8™ Health Survey norms are based on a 2000 U.S. general population sample. Additionally, as previously described, several sets of international norms are available for use with the SF-36® Health Survey. Although international norms for the SF-36v2™ Health Survey are not as abundant as those for its predecessor, the number of SF-36v2™ Health Survey translations is continually growing.

***Norm-based scoring and interpretation.***

The desire for norm-based scoring (NBS) and interpretation guidelines no longer needs to be a consideration when choosing among the Short Form surveys. NBS can be used to score all Short Forms (see Chapter 14 for detailed information).

***Availability of health domain scales.***

Interest in the eight health domains, in addition to the two component summary measures, is no longer a reason for favoring the two SF-36® Health Surveys over the two SF-12® Health Surveys. In contrast to the SF-12® Health Survey, which yielded score estimates for only the two component summary measures (Ware, Kosinski, & Keller, 1995, 1996), the SF-12v2™ Health Survey has the advantage of yielding scores for all eight health domains and the physical and mental component summary measures. The SF-8™ Health Survey also provides scores on all health domain scales and component summary measures.

***Translations.*** Both versions of the SF-36® Health Survey, both versions of the SF-12® Health Survey, and the SF-8™ Health Survey have been translated or adapted into a total of more than 60 languages and other translation projects are currently underway. Issues and considerations regarding translated versions of the SF instruments are

discussed later in this chapter. A list of translated versions of all Short Form instruments is available at <http://www.qualitymetric.com/products/license/AboutLicensing.aspx#>.

***Documentation.*** Up-to-date manuals documenting survey development, scoring algorithms, U.S. general population norms, and interpretation guidelines are available for all adult Short Form instruments.

***Published literature.*** By August 2006, over 8,500 articles and other publications about the Short Form surveys were identified. Most of these publications (more than 7,000) are about the SF-36® Health Survey. This may be an important consideration in instrument selection if the objective of a survey or study is narrow in focus and benchmarks from the published literature are crucial. With the noteworthy improvements achieved with the SF-36v2™ Health Survey, the number of published articles on this version of the survey is expected to accelerate quickly within the next few years. Updated information on all of the Short Form surveys is available at <http://www.qualitymetric.com> and <http://www.sf-36.org>.

## **Matching a Form to an Application: General Considerations**

Because of improvements incorporated into the SF-36v2™ Health Survey and SF-12v2™ Health Survey, these updated surveys are recommended over their original versions. The updated surveys are frequently considered the tools of choice for fixed-form, short form questionnaires and are recommended for use in clinical trials, outcomes and effectiveness research, and clinical practice applications. Aside from these situations, a number of factors should be considered when deciding which survey to use for a particular application. The decision hinges in large part on making a tradeoff between respondent burden and score precision. This and other considerations are addressed below.

***Assessing and monitoring individual patients for clinical purposes.*** Originally, the SF-36® Health Survey was used in population health surveys. Its brevity, however, has made it and the SF-36v2™ Health Survey increasingly attractive for use in clinical trials and for individual patient evaluation purposes in clinical practice.

Selection from among the available health status measures for the assessment and monitoring of individual patients for clinical purposes often represents a compromise between the burden that is placed on the patient and medical staff to obtain the information and the usefulness of that information. Obtaining health domain and component summary information is much less burdensome when employing the SF-12v2™ Health Survey instead of the SF-36v2™ Health Survey, and even less burdensome when using the SF-8™ Health Survey. At the same time, the SF-12v2™ Health Survey and SF-8™ Health Survey cover a narrower range of functioning and are less precise than the SF-36v2™ Health Survey (see Table 3.1). Thus, the two shorter instruments provide less quantitative and reliable information about a patient's health status at any given point in time and the amount of change in that status over time. Therefore, use of the SF-12v2™ Health Survey or SF-8™ Health Survey for assessing and/or monitoring individuals is discouraged. Instead, DYNHA® Computer Adaptive Health Assessments are recommended for this purpose, unless a fixed-form instrument is required, in which case the SF-36v2™ Health Survey is recommended. Use of the SF-36v2™ Health Survey provides greater utility and breadth of coverage at both the component summary measure and health domain scale levels. For example, its five-item MH scale, initially developed as the Mental Health Inventory (MHI-5; Berwick et al., 1991; Veit & Ware, 1983), has been found to be a psychometrically sound alternative to longer instruments for the screening of anxiety and affective disorders (Berwick et al., 1991). Its

usefulness with individual patient evaluations has also been established in case study demonstrations (e.g., see Wetzler, Lum, & Bush, 2000; see also Chapter 12).

It is important to note that some experts in the field would contend that the psychometric properties of the SF-36v2™ Health Survey are not adequate for use in individual assessments. For example, McHorney and Tarlov (1995) argued that the SF-36® Health Survey did not meet all of their six criteria for individual patient applications. These criteria were: (a) practical features (e.g., takes less than 15 minutes to complete), (b) breadth of health measured (e.g., includes scales for measuring physical and mental status), (c) depth of health measured (e.g., allows for adequate floor and ceiling), (d) cross-sectional measurement precision (e.g., internal consistency reliability greater than or equal to .90), (e) longitudinal-monitoring measurement precision (e.g., 2- to 4-week test-retest reliability greater than or equal to .90), and (f) validity (e.g., convergent and divergent validity, sensitivity to change).

According to the data available at the time, McHorney and Tarlov (1995) argued that the SF-36® Health Survey did not meet the above stated criteria for ceiling effects and reliability (internal consistency and test-retest). However, these requirements may be too stringent and unrealistic. By these standards, the Minnesota Multiphasic Personality Inventory-2 (MMPI-2; see Butcher et al., 1989), arguably the most widely used and researched objective personality assessment tool in the world, would not be considered appropriate for individual testing purposes because of the reliability of its scales (Butcher et al., 1989, Tables D-1 through D-9). Regarding ceiling and floor effects, the floor effects of the SF-36® Health Survey health domain scales for which this was particularly problematic, RP and RE, were significantly reduced when revised for the SF-36v2™ Health Survey.

Furthermore, the required “practical features” can realistically come only with

some sacrifice in other required features, whether it is lowered validity or reliability or limitations in the breadth or depth of measurement. In some cases, as with the SF-36v2™ Health Survey MH scale mentioned above, brevity may not require such a compromise. In short, many experts would argue that the SF-36v2™ Health Survey is more than “adequate” or “acceptable” for individual patient assessment, especially in light of the demands that healthcare systems place on such instruments (e.g., brevity, ease of use) if they are to be incorporated into the daily work flow of care providers (e.g., Maruish, 2002).

Perhaps more importantly, the provider considering using the SF-36v2™ Health Survey must decide whether an evaluation of a patient is better served with or without the information that it provides. It is the contention of its developers that SF-36v2™ Health Survey results for an individual patient will always contribute to the evaluation of that patient by providing either new information or information that supports or clarifies the provider’s clinical impressions. Further discussion on its use for clinical purposes can be found in Chapter 2 and is illustrated in Chapter 12.

**Detecting small group differences.** A high standard of score reliability (.90 or higher) is recommended to achieve satisfactory statistical power, and single-item health scales like those in the SF-8™ Health Survey are likely to be inadequate or unable to detect only very large differences. In these situations, use of CAT and the DYNHA® Computer Adaptive Health Assessments would provide the best solution. However, the SF-36v2™ Health Survey and SF-12v2™ Health Survey are recommended for efforts focused on detecting small group differences when the DYNHA® Computer Adaptive Health Assessments is not an administration option. The improved precision afforded by the two longer measures can be observed through narrower confidence intervals around score

estimates.

**Large population surveys and samples.** The SF-36v2™ Health Survey, SF-12v2™ Health Survey, or SF-8™ Health Survey may each be considered for use in the largest population surveys and for studies involving large samples and group-level comparisons. Single-item measures, such as those used for all scales in the SF-8™ Health Survey and four of the eight SF-12v2™ Health Survey scales, work well in these situations because the precision of mean scores is determined more by sample size than by increasing measurement reliability. Although concerns have been expressed in the past about single-item measures, several of these concerns are addressed by the use of norm-based scoring algorithms (see Ware, Kosinski, Dewey, & Gandek, 2001), making the SF-8™ Health Survey an appropriate choice for large surveys of representative samples. Furthermore, because statistical power is in part a function of sample size, the SF-8™ Health Survey may be a more viable and practical tool to use in large population studies.

**Ongoing studies.** The authors recommend against adopting either the SF-36v2™ Health Survey or SF-12v2™ Health Survey in “midstream;” that is, during the course of a longitudinal study that began by using the SF-36® Health Survey or SF-12® Health Survey, respectively. Unless the number of years remaining in a longitudinal panel study is large, the threat to validity and the reasons for concerns perceived by others may be too great to justify the change. In such cases, parallel administrations of items from the two versions of the SF-36® Health Survey or the SF-12® Health Survey may provide the additional data necessary to determine whether estimates of scores generalize across the two versions of the instrument. With the availability of 1998 NBS algorithms for both versions of the SF-12® Health Survey and both versions of the SF-36® Health Survey, there is now the link

required for meaningful comparisons of results between the two versions of each survey.

**Cross-cultural studies.** One of the important features of the Short Form surveys is the availability of translated versions for use in non-English speaking countries or with U.S. samples in which English is not the first or primary language. Translations are available for each of the five Short Form surveys, with the SF-36® Health Survey offering the greatest number of validity and normative studies to date (for example, see Gandek & Ware, 1998b). The focus now, however, is on developing more validated translations for the SF-36v2™ Health Survey, SF-12v2™ Health Survey, and SF-8™ Health Survey. Users requiring a translated version of one of the Short Form surveys can consult the SF-36® Health Survey Web site (<http://www.sf-36.org/>) or the QualityMetric Incorporated Web site (<http://www.qualitymetric.com/>) for a current list of available translated versions for each instrument. Both Web sites also provide links to the International Quality of Life Assessment (IQOLA) Project Web site. SF users should contact QualityMetric Incorporated for recommendations if a desired translation for a specific Short Form is not available.

Table 3.3 summarizes some of the general similarities and differences among the Short Form surveys.

### **Matching a Form to an Application: Specific Form-to-Form Considerations**

**SF-36® Health Survey versus SF-36v2™ Health Survey.** The SF-36v2™ Health Survey is recommended over the SF-36® Health Survey for all *new* studies requiring the administration of a Short Form survey instrument, including population surveys, outcomes research studies, and controlled clinical trials, as well as for research studies and applications in clinical practice focusing on results of individual patients. For all of these applications, the SF-36v2™ Health

Survey is superior to its predecessor in a number of respects, as noted previously.

Comparability of results and the availability of interpretation guidelines are important considerations in choosing a health status measure. The NBS algorithms and 1998 norms documented in Chapter 14 make it easy to interpret SF-36v2™ Health Survey results and also make it possible to compare these results to those for obtained with the SF-36® Health Survey. NBS and the 1998 norms provide the link between the two versions, while making both forms easier to interpret in relation to population norms. Users of the SF-36® Health Survey will find the 1998 norms more up to date and NBS scores for the eight health domain scales easier to interpret; for the same reasons, NBS makes the SF-36® Health Survey component summary measures easier to interpret. These same considerations and recommendations apply to the use of the SF-12v2™ Health Survey over the SF-36® Health Survey.

**SF-36v2™ Health Survey versus SF-12v2™ Health Survey.** The SF-12v2™ Health Survey is the instrument of choice in surveys that require a shorter instrument than the SF-36v2™ Health Survey. Large population health surveys can take advantage of its brevity (in comparison with the SF-36v2™ Health Survey) with confidence that, with only rare exceptions, group differences and changes in health status over time will be detected and that scores and interpretive guidelines will be directly comparable with those from the SF-36v2™ Health Survey. The fact that the SF-12v2™ Health Survey is a subset of 12 items from the SF-36v2™ Health Survey is a noteworthy advantage if the objective is maximum comparability of results and equivalence of population norms and other interpretation guidelines developed for the longer instrument. Most publications of “head-to-head” comparisons between the SF-12® Health Survey and SF-36® Health Survey, including studies of responsiveness,

reach the same conclusions about the PCS and MCS measures (see Ware, Kosinski, Turner-Bowker, & Gandek, 2002). Among the most common criticisms noted in published reports from those studies are the observed ceiling and floor effects, particularly for the two SF-12<sup>®</sup> Health Survey role participation scales. However, the developers did not intend for the eight health domain scales to be scored from SF-12<sup>®</sup> Health Survey item responses because of their coarseness and observed ceiling and floor effects. The SF-12v2<sup>™</sup> Health Survey represents a substantial improvement in that regard (see Table 3.1) and provides a means of scoring the health domain scales as well as the PCS and MCS measures.

**SF-12v2<sup>™</sup> Health Survey versus SF-8<sup>™</sup> Health Survey.** The SF-8<sup>™</sup> Health Survey provides an even shorter option for purposes of estimating the health domain scale and component summary measure scores in the largest population health surveys. However, unlike the SF-12v2<sup>™</sup> Health Survey, items in the SF-8<sup>™</sup> Health Survey are not a subset of those in the SF-36v2<sup>™</sup> Health Survey, and this may be a disadvantage depending on the purpose of the study and the degree of direct comparability demanded (see Ware, Kosinski, Dewey, & Gandek, 2001). Scores for all eight health domains are estimated from single-item SF-8<sup>™</sup> Health Survey measures, as are scores for four of the eight SF-12v2<sup>™</sup> Health Survey scales. As noted earlier, such single-item measures work best in very large surveys of general and specific populations in which precision is achieved much more by drawing upon a large representative sample than by increasing measurement reliability. The SF-12v2<sup>™</sup> Health Survey is also the instrument of choice for surveys that need greater precision over a wider range of levels of health than can be measured using the SF-8<sup>™</sup> Health Survey.

Concerns about single-item measures still apply (McHorney, Ware, Rogers, Raczek, & Lu, 1992; Ware, Kosinski, & Keller, 1996); however, concern has

diminished due to advances in item response categories and improvements in scoring algorithms for single-item scales. Also, there is a better understanding of the conditions under which the standard error of the measurement of an *individual*, as opposed to the standard error of a *group mean*, is and is not worth a substantial increase in respondent burden. The usefulness of well-constructed, single-item measures in group-level clinical trials and outcomes research projects is a subject of considerable interest and research (e.g., Aoki, Fleming, Griffin, Lacy, & Edmundson, 2000; Paterson, Langan, McKaig, et al., 2000; Silagy, Griffin, Lacey, & Edmundson, 1998; Ware, Kosinski, Dewey, et al., 2001).

**Short Form fixed-form measures versus CAT.** Seeking the highest level of accuracy may be required for those survey applications focusing on individual scale scores or needing to detect the smallest of important changes in health status in very small group-level analyses. For the most demanding applications, one no longer needs to rely on short or long fixed-form instruments to achieve more practical or more precise measures. Research in progress suggests that software based on CAT logic provides the best solution.