Chapter 7
Interpretation of the RAND–36 HSI Scores

Uses of measures of overall health status similar to the RAND–36 HSI have been based largely on aggregate-level analysis. These uses have included the monitoring of population health (Aaronson et al., 1992; McHorney, Kosinski, & Ware, 1994); estimating the burden of different conditions (Hays, Wells, Sherbourne, Rogers, & Spritzer, 1995; McHorney et al., 1993; Sherbourne et al., 1994); and clinical trials of treatment effects and the monitoring of outcomes in clinical practice, such as quality-assurance evaluation in hospital-based outpatient clinics. For studies involving clinical trials of treatment effects, traditional clinical parameters have been used more than have generic health status measures. However, an evaluation of general functioning when equally efficacious treatments are being compared affords an evaluation of these treatments with respect to their different trade-offs in general functioning and well-being (Fowler et al., 1988). Some researchers have recommended the combined use of general health measures, such as the RAND–36 HSI, and disease-targeted measures, for more breadth and depth in assessing outcomes (Patrick & Deyo, 1989).

As the need for the clinical application of health status instruments has increased, attention has focused on the clinical sensitivity of the instrument. For instance, research has been conducted to determine the minimum size of a group necessary for detecting clinically meaningful differences or changes and for determining the smallest mean group differences in health scale scores that would be considered clinically and socially relevant.

The increased need for instruments that can reliably detect significant change has led to a shift from individual scales, such as those on the SF–36, to composite scores because of the increased reliability and statistical sensitivity. Simply put, the change in scores on an individual scale with a few items must be much greater to be statistically significant due to lower reliability than a change in composite scores, which have more levels of difficulty and higher reliability.

Individual-level interpretations, although less frequent than aggregate-level analysis, have been based on the following approaches:

- With content-referenced interpretation, an individual's scores are referenced to the specific content of different response levels and to the percentage of individuals from a normative sample who have responded at these response levels.

- With criterion-based interpretation, an individual's scores are related to the percentage of people who report a specific criterion, such as inability to work or frequency of utilization of various services. The average amount of change reported in medical outcome studies may be cited.
- With norm-referenced interpretation, an individual's scores are compared to mean scores and designated percentile rankings for various demographic and disease-specific groups.

Determination of Clinical Meaningfulness

General Parameters

Individual-level interpretation of RAND–36 HSI performance relies primarily on the three composite scores. The Physical Health and Mental Health composites represent the physical and mental aspects of functioning and well-being consistent with the identification of predominantly physical or predominantly mental disorders. The Global Health Composite provides a global measure that encompasses overlapping aspects of physical health and mental health for use as a more integrated assessment of health-related limitation. In addition, discrepancies between the Physical Health and Mental Health composite scores can be interpreted. The tracking of client change over time based on the scale scores is not recommended because of insufficient reliability. However, the scale scores can be used to further describe an individual's composite scores at specific points in time.

Missing Item Responses

With the SF–36, a scale score could be computed as long as at least half the items were completed on each scale. This practice, which substitutes estimates for missing data, has been found to be too imprecise for individual profile analysis. For the RAND–36 HSI, only one item per scale may have a missing response, and only three items may have missing responses for the entire inventory. These restrictions maximize individual-level accuracy. Obviously, then, individuals should be encouraged to complete all items. (Computation of estimated IRT weights for missing item responses is described in Appendix A.)

Clinical Cut Scores

The clinical meaningfulness of the scores obtained on the RAND–36 HSI is determined by the integration of three sources of information—*T* scores, cumulative percentages of *T* scores, and criterion-based cut scores. As described previously, *T* scores are a way of standardizing raw scores on scales for each normative group in a way that allows comparability across normative groups. For example, although individuals of different ages may obtain different scale raw scores, their *T* scores may be the same when based on the norms for their own age cohorts. Also, placing the scale scores on the same *T*-score metric ($M = 50$, $SD = 10$) allows comparability across scales that have different raw score means and standard deviations.

On the other hand, individual-level clinical interpretation in terms of *T* scores may be difficult when distributions are strongly skewed. Although 1 *SD* below the mean score of a scale represents a significantly lower score in terms of overall variability of the scale, its clinical meaning may vary. This variation occurs because the same linear *T* score reflects different cumulative percentages across scales and even across the seven normative groups on the same scale. Also, because of the skewed nature of these scales, the mean *T* score (50) does not necessarily represent the median point, or the *T* score obtained by 50% of the individuals in the normative sample. For these reasons, a cumulative percentage is presented for each *T* score and should be considered along with the *T* score in any clinical interpretation.

Although *T* scores and cumulative percentages provide information about an individual's relative standing in his or her normative group with respect to variance and frequency with which the score occurs, they do not provide information about independently measured health conditions. Previous research has attempted to provide anchor points by providing raw score and *T*-score means of groups of individuals manifesting specific diseases. This approach is helpful particularly if the individual has manifested that particular disease and has demographic characteristics similar to those of the reported group. It should be kept in mind, however, that general perceived-health-status measures are designed to cut across specific disease categories. For this reason, cut scores for the RAND–36 HSI are based on criteria that are not disease-specific but that pertain in a more general way to the dimensions assessed by each composite. The development of local norms by researchers working with disease-specific populations, however, is encouraged.

The guidelines presented in this chapter for evaluating an individual's scores integrate the use of *T* scores, cumulative percentages, and cut scores derived from discriminant function analyses, which were based on the scores obtained by the age-stratified normative group to ensure that results were demographically reflective of the U.S. population. It is emphasized that the cut scores are intended only as guidelines to be carefully considered in the context of local norms, disease-specific information, and circumstances unique to the individual. The following principles were used for determining these cut scores:

- *Low* scores are defined as *T* scores obtained by approximately 20% or less of the normative group.
- The *clinical cut score* is defined as the point that optimally differentiates those in an independently defined criterion group as *low* on the criterion from those defined as *high* on the criterion (see discussion in Chapter 5). The low-criterion group is actually composed of those respondents who obtained *high* scores on the criterion measures of symptom or dysfunction, and the high-criterion group is actually composed of those who obtained *low* scores on those criterion measures. Those *T* scores below the clinical cut score are defined as *below criterion;* those *T* scores above this point are defined as *above criterion.* For the Physical Health Composite, those reporting more than one physical condition, one or more of which were self-reported as limiting functioning to some degree, were compared with those who reported the presence of no limiting physical conditions. For the Mental Health Composite, those who reported moderate to severe psychological symptoms of depression and anxiety were compared with those who reported few or none of these symptoms. For the Global Health Composite, those reporting difficulty in life functioning were compared with those reporting no difficulty in life functioning.
- *High T* scores are defined as those >50 and obtained by approximately 50% or more of the normative nonclinical sample.

Sequential Analysis

The recommended logic for clinical use of the RAND–36 HSI scores is sequential, proceeding from general to specific. This logic is supported by factor analysis, which indicates that health status is an underlying integration of two predominant dimensions, physical and mental, and that these dimensions are, in turn, composed of specific aspects. This logic is also supported by the psychometric reality that the composite scores have higher reliability than the scale scores and, therefore, are more statistically appropriate for individual interpretation. The interpretive guidelines provided here recommend the consideration of individual scale scores only in the context of the statistically stronger composite scores. The following steps for the clinical interpretation of an individual's scores proceed from the general to the more specific.

Suggested Guidelines

Physical Health Composite

Low Scores

A *T* score of 42 or lower was obtained by 19.8% of the age-stratified normative sample. Such low scores, therefore, are relatively infrequent in a nonclinical sample. In addition, a subgroup of the normative sample (*n* = 124) who had reported diagnosed disabilities obtained a mean *T* score of 39.9. These findings support the conclusion that *T* scores ≤42 suggest that perceived physical health problems are impeding life functioning.

Clinical Cut Score

A *T* score of 47 was determined by discriminant function analyses (see Chapter 5) to be the cut score below which 60.5% of those reporting physical conditions were accurately identified. A *T* score >47 accurately identified 89.5% of those not reporting any physical conditions. A *T* score >47 is higher than the scores obtained by 29% of the age-stratified normative sample. This cut score yielded 39.5% false-negatives (high-symptom respondents who obtained *T* scores >47) and 10.5% false-positives (low-symptom respondents who obtained *T* scores ≤47).

High Scores

A *T* score of 53 or lower was obtained by 54.2% of the age-stratified normative sample. Scores >53 suggest that these individuals are less likely to have physical health problems that impede life functioning.

Mental Health Composite

Low Scores

A *T* score of 38 or lower was obtained by 14.4% or less of the age-stratified normative sample. Results of discriminant function analyses indicated that a *T* score of 38 accurately identified 66.7% of those reporting moderate or severe symptoms of depression and/or anxiety and 96.5% of those reporting minimal symptoms. A *T* score of 38 yielded 33.3% false-negatives and 3.5% false-positives. Therefore, a *T* score of 38 or lower was likely to indicate an individual in the criterion group who was reporting psychological symptoms that might impede life functioning.

Clinical Cut Score

A cut *T* score ≤41 was associated with accurate identification of 56% of those in the age-stratified normative sample who reported from mild to severe depression or anxiety. A *T* score ≥42 was obtained by 93.8% of those who reported minimal symptoms of depression or anxiety. Also, a *T* score ≤41 on the Mental Health Composite was obtained by only 18.6% of the age-stratified sample. This cut score of 41 yielded 44% false-negatives (high-symptom respondents obtaining *T* scores >41) and 6.2% false-positives (low-symptom respondents obtaining *T* scores <42) in predicting scores of depression and anxiety.

High Scores

A *T* score of 53 or lower was obtained by 53.2% of the age-stratified normative sample. Individuals obtaining scores >53 are not likely to perceive mental health problems that impede life functioning.

Physical Health and Mental Health Composite Score Discrepancy

The *T* scores obtained by an individual on the Physical Health and Mental Health composites should be compared to determine their similarity or dissimilarity. The clinical meaning of a discrepancy, although not yet explored empirically, may be reviewed for each respondent in the context of his or her unique circumstances. A discrepancy might indicate that there is a perceived difference between the person's physical health and mental health. Minimally, a significant discrepancy would raise questions about the use of the Global Health Composite score as a single indicator. The following formula is used to calculate the amount of discrepancy required for statistical significance:

$$D = z\sqrt{SE_{M_a}{}^2 + SE_{M_b}{}^2}$$

where D is the difference between the two composite scores, $z = 1.64$ at the 90% level of confidence, and $SE_M = SD\sqrt{1 - r}$, where SD is the standard deviation and r is the reliability coefficient of the respective composite score.

For the age-stratified sample, a discrepancy between the Physical Health and Mental Health composite scores greater than or equal to ±6.35 would be considered statistically significant at the 90% level of significance. However, the clinical significance of a discrepancy may be suggested more appropriately by the infrequency of the discrepancy. For this reason, a discrepancy greater than or equal to ±10 is recommended because less than 22% of the age-stratified sample obtained this discrepancy. This amount also represents 1 *SD* difference between the two composite scores. Appendix D provides tables showing the cumulative percentages of individuals in all seven normative groups obtaining various discrepancies between these composite scores. Table D.1 provides this information for those whose Physical Health Composite score exceeded the Mental Health Composite score; Table D.2 provides the information for those whose Mental Health Composite score exceeded the Physical Health Composite score.

Global Health Composite

The Global Health Composite score reflects the overall perceived health status of the individual and may be viewed as a "global thermometer" of the individual's well-being. The meaning of the Global Health Composite score is more clearly interpretable when the Physical Health and Mental Health composite scores are not significantly discrepant.

Low Scores

A T score of 42 or lower was obtained by 21.0% of the age-stratified sample. Such low scores may be considered relatively infrequent and as suggesting that perceived health problems are impeding life functioning.

Clinical Cut Score

Discriminant function analyses (see Chapter 5) identified a cut score of 50, which accurately identified 80.5% of *low functioners* obtaining a *T* score <50 and 92.4% of *high functioners* obtaining *T* scores ≥50. Low functioners were defined as those who were among the highest 25% of scorers on the SAS–SR (Weissman & Bothwell, 1976) Global Scale and the BASIS–32 (Eisen et al., 1994) Daily Living/Role Functioning Scale. High functioners were those who obtained among the lowest 25% of scores on these scales. A *T* score of ≤49 was reported by 37% of the RAND–36 HSI age-stratified normative sample. This cut score yielded 19.5% false-negatives (low functioners who obtained *T* scores ≥50) and 7.6% false-positives (high functioners who obtained *T* scores ≤49).

High Scores

A *T* score of 52 or lower was obtained by a least 50.8% of the age-stratified sample. Scores >52 may be considered sufficiently prevalent in the nonclinical population to indicate normal or better global health functioning.

Scale Scores

Caution should be exercised in interpreting scale scores in isolation. Scale score patterns may be productively considered as interview cues to the extent that the pattern presents clinically relevant questions. For example, "Your answers seem to suggest that while you feel a good deal of pain, it has not limited your daily activities too much. Is that accurate?"

A Final Caution

It should be noted that the recommended cut scores vary across composites. This variance occurs because the composite scores are associated with different cumulative percentage distributions and because different independent measures were used to define criterion groups. The Physical Health and Mental Health cut scores were more symptom-based, and the Global Health cut score was related to self-reported functioning. Therefore, it is possible for an individual's Physical Health and Mental Health composite scores to reflect a minimal likelihood of physical or mental problems and for the Global Health Composite score not to be in the corresponding range of global health status. Because an individual may be relatively symptom-free physically and mentally but still not enjoy good health or be functioning adequately, clinical judgment should be exercised in interpreting all scores.

Longitudinal Tracking of Change

Objective assessment to accurately determine if an individual has benefited from clinical intervention is essential. One method of assessing change that has proved useful is the *Reliable Change Index* developed by Jacobson, Follette, and Revenstorf (1984) and further described by Jacobson and Truax (1991). Briefly, with this approach, the extent of an individual's change in health status is assessed by determining if he or she has approached a normal population's responses and diverged farther from the response pattern of a dysfunctional population.

The method of assessing meaningful clinical change described for the RAND–36 HSI is related to this method but differs in important ways. The analysis recommended here is a two-step process. First, a determination is made of whether the difference in an individual's scores between one point and the next is statistically different. If the difference is statistically significant and positive, then the next step in interpretation is taken. In Step 2, the change is interpreted as *positive but insufficient, favorable, very favorable,* or *optimal.* These clinical anchors are based on the cut-score criteria provided earlier.

The three composite scores should be used for tracking clinical change because these scores are psychometrically reliable enough to allow differentiation of significant change from random fluctuation. The scale scores, on the other hand, do not consistently provide this psychometric reliability. For example, the range of reliability coefficients for the scales included in the Mental Health Composite is from .71 to .84. These values would be associated with 90% confidence intervals ranging from ±8.90 to as much as ±11.55 (on the *T*-score metric). In contrast, the confidence intervals for the Mental Health Composite range from ±5.60 to ±7.79.

The composite scores chosen for comparison are usually sequentially ordered; that is, an earlier score is compared with a more recent one in order to determine whether intervening factors, such as a clinical intervention, might have affected the individual's health status in some way. It is important for clinicians to keep in mind that a change in health status may be attributed to many things, including the course of a particular disease or the recovery process.

The clinician may choose an optimal time interval between administrations of the RAND–36 HSI. This time interval should take into account the nature of the individual's clinical condition and realistic expectations about the course of change. The clinician should also ensure that the normative group used to determine an individual's *T* scores at different times of testing is the same. The following guidelines are recommended.

- Using age-specific norms is preferred for evaluating perceived health status with the RAND–36 HSI because perceived health status is most sensitive to changes in age. For this reason, short-term comparison of an individual's health status should be based on the *T* scores for the individual's specific normative age group. A composite raw score for someone aged 18–24 will mean something quite different from the same raw score for an individual over 65 years old. On the other hand, if a longer time period has elapsed between test administrations, so that an individual shifts from one normative age group to an older one, the individual's *T* scores should be based on the specific age-group norms appropriate at each time of testing. In this way, the individual's perceived health status will be evaluated relative to what is normal for his or her age cohort at the time of testing. Moreover, an individual's perceived health status over the course of a lifetime may be evaluated relative to age. Of course, *T* scores based on age-specific norms should not be compared to *T* scores based on overall or gender-specific norms.
- On the other hand, there may be occasions when the evaluation of an individual's perceived health status should not be controlled for age so that difference due to age-related changes are discernible. In such cases, *T* scores should be based on the overall,

male, or female normative group. Moreover, the same normative group (e.g., female) should be used for all longitudinal comparisons of an individual's status.

Step 1: Determining Statistical Significance of Change

The statistical significance of a change in scores is based on the individual's estimated true score and the confidence interval of that score, which is determined by the standard error of prediction. The estimated true score (ETS) is calculated with the following formula, which corrects for regression to the mean:

$$\mathrm{ETS} = M + r(x - M),$$

where *M* is the mean (50), *r* is the reliability coefficient of the score, and *x* is the observed test score. Because the composite scores used in longitudinal tracking are *T* scores, the mean for all samples is 50 and the standard deviation is 10. The reliability coefficients of the scores are those provided in Table 5.1.

The estimated range of anticipated fluctuation in scores from first to second testings is based on the standard error of prediction (*SEp*), which is calculated with the following formula:

$$SE_p = (1.64)\ SD\sqrt{1 - r^2},$$

where 1.64 is the *z* value at the 90% level of confidence, *SD* is the standard deviation (10), and *r* is the reliability coefficient of the score.

The standard error of prediction is used to establish the confidence interval around the estimated true score from the first testing. Its value is subtracted from and added to the estimated true score. The resulting values are the lower and upper ends, respectively, of the confidence interval. The standard errors of prediction for the Physical Health, Mental Health, and Global Health composite *T* scores for all seven normative groups are provided in Table 7.1.

If the individual's composite score from the second testing falls within the confidence interval established for the estimated true score from the first testing, then change is rated as *equivocal.* An equivocal rating means that there has been no statistically significant change in the composite scores. If the score from the second testing is above the confidence interval, that is, greater than the highest score in the range of scores, the change is rated as *positive.* A positive rating indicates that there has been a statistically reliable increase in the composite score, reflecting some improvement in health status. If the score from the second testing is below the confidence interval, then there has been a statistically reliable decrease in the individual's health status as measured by the composite score, and progress is rated as *negative.*

Table 7.1. Standard Errors of Prediction for the Three RAND–36 HSI Composites at 90% Level of Confidence

Normative Group	Physical Health Composite	Mental Health Composite	Global Health Composite
Age-Based Sample			
18–24	±6.80	±7.79	±6.03
25–44	±5.60	±6.80	±5.12
45–64	±5.12	±5.60	±4.59
≥65	±5.60	±7.79	±5.60
Age-Stratified Sample			
Overall	±5.60	±6.80	±5.12
Female	±5.60	±6.43	±5.12
Male	±5.60	±7.48	±5.12

Briefly, the statistical significance of a change in composite scores is determined by

- calculating the individual's estimated true score for the first testing;
- establishing the confidence interval of that score with the standard errors of prediction provided in Table 7.1; and
- rating the change in scores as *equivocal, negative,* or *positive* according to the second score's position relative to the confidence interval.

If an individual's composite score comparison indicates an equivocal (or no) change, the person's condition is described as unchanged from previous assessment, and analysis is discontinued until future testing. In such cases, the clinician should keep in mind that change is sometimes so gradual that it may not be apparent if the time between testings is not sufficiently long for change to have occurred. If change has not occurred or is in a negative direction, the course of treatment should be evaluated and clinical decisions made appropriate to the specific case and appropriate treatment guidelines. If the comparison reveals a significant improvement in health status, then the clinical meaningfulness of the change is evaluated.

Step 2: Evaluating the Clinical Meaningfulness of Change

If the analysis in Step 1 revealed significant change in a positive direction, a subsequent composite *T* score of the individual is addressed in terms of relative health desirability and anticipated goals of possible interventions. Table 7.2 lists the *T*-score ranges associated with determining if a positive change is *positive but insufficient, favorable, very favorable,* or *optimal.* These *T*-score ranges are based on the empirical guidelines suggested earlier in this chapter. Thus, improvement would be rated as *optimal* if the individual's subsequent Global Health Composite *T* score is significantly greater than his or her previous *T* score and greater than or equal to 52.

Table 7.2. ***T*-Score Ranges for Evaluating the Clinical Meaningfulness of Change**

	Positive but Insufficient	Favorable	Very Favorable	Optimal
Physical Health Composite	≤42	43–46	47–52	≥53
Mental Health Composite	≤38	39–41	42–52	≥53
Global Health Composite	≤42	43–49	50–51	≥52

Summary

Finally, the interpretation of scores and related clinical decisions will depend on the unique clinical circumstances of the individual respondent and the prevailing practice guidelines. For instance, it should be noted that an evaluation of significant change is made independently for each composite because cut scores were based on separate score distributions and criterion groups. It is possible for an individual's improvement in mental health status to be evaluated as favorable but for his or her change in global health not to reflect a similarly favorable status. The clinician must use his or her clinical judgment and consider all known aspects of the individual's circumstances when interpreting the individual's scores.