Chapter 3
Item Scaling and Development of the Composites

Item Scaling

Theoretical Background

The original *RAND 36–Item Health Survey* (Hays et al., 1993) and 33 of the 36 items on the SF–36 utilize a traditional method of scoring, namely, a simple summation approach. However, this technique is based on certain implicit assumptions about the items. The first assumption is that each item has options that fit an interval scale. In other words, the options are assumed to be equidistant in terms of the metric of the construct underlying the scale. For example, if the three response options for an item are *not at all*, *somewhat*, and *very much*, then the assumption is that the difference in the construct being measured between *not at all* and *somewhat* is the same as that between *somewhat* and *very much*.

The second assumption in the simple summation approach is that all of the items should contribute equally to the overall scale. This assumption means that a response of *somewhat* represents the same amount of the underlying construct for every item in a scale. Further, there is the implicit assumption that the overall scores on a scale are based on an interval scale. For example, the difference between scores of 7 and 8 is the same as that between scores of 8 and 9. The difficulty presented by such an approach is that on a scale of illness with the first 8 items indicative of very minor illness and the last 2 indicative of very severe illness, a person who answered the first 7 items affirmatively is actually only slightly less ill than one who answered the first 8 items affirmatively. However, the difference in illness indicated by these two individuals and by someone who endorses the first 9 items is considerably larger.

In item response theory (IRT), the relationship between examinee item performance and the latent trait can be described by an item characteristic function that is monotonically increasing (Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991; Hays, 1998). The advantage of using latent trait estimation from IRT is that an estimate of how much each response should contribute to the overall score can be made and depends on the underlying level of the construct associated with that item response. Thus, on a measure of psychological distress, a response of "sometimes" to "I feel that I am going to die" would be

assigned a higher score (on this continuum) than a response of "often" to "I feel uncomfortable." With IRT, the assignment of differential scores to responses is possible because all of the responses, across and within items, can be placed on a latent-trait continuum.

Scoring for the RAND–36 HSI scales is based on a one-parameter IRT model (1-PL). According to this IRT model, the expected score of a respondent on a particular item is a function of both the item difficulty and the ability (latent trait) of the respondent (Rasch, 1960, 1966). Within the 1-PL family, the Rasch model was used for scales with all dichotomous-response items (e.g., 1 or 0 or yes/no) and the partial credit model was used for scales composed of items with multicategory response options (Masters, 1982). A different weight is assigned to each response option within items of the same scale by placing all response options of the same scale on the same underlying latent-trait continuum.

Previous Research

With earlier applications of the SF–36 scales, a simple summation approach for scoring was used. The assumption was that simple scoring was possible because items of the same scale had roughly equivalent relationships to the underlying health construct being measured. Researchers initially suggested that it was not necessary to standardize or weight the SF–36 items (McHorney, Ware, et al., 1994; Ware, Snow, Kosinski, & Gandek, 1993). In examining the results from 24 different patient and demographically diverse groups, McHorney, Ware, et al. (1994) also maintained that simple summation was warranted on the basis of Likert's (1932) assumptions of similar item means, score variances, and item–total correlations.

More recently, evidence contrary to Likert assumptions has been presented. Ware et al. (1993) suggested that of the SF–36 items, two were shown by empirical data not to satisfy the assumptions of a linear relationship between item scores and the underlying health construct defined by the scale. According to Keller, Ware, and Gandek (1995), observed departures from equal-interval assumptions were consistent across countries and were the greatest for the two response scales that had been recalibrated in the SF–36 scale scores (*excellent, very good, good, fair, poor* and *none, very mild, mild, moderate, severe, very severe*). One item of the Pain Scale was found to have severity ratings that did not satisfy the assumptions of equal intervals. Response-option values were consequently recalibrated for the SF–36 and values derived from the mean values of a summary criterion; the values computed were the mean value for respondents who chose each of the six levels defined in Item 7 of the same scale (Ware et al., 1993).

Recalibration had been recommended for the item measuring general health on the General Health Perceptions Scale (Davies & Ware, 1981; Stewart et al., 1988; Ware, Nelson, Sherbourne, & Stewart, 1992). These researchers found that the mean value for a criterion of general health for the respondents who chose each of the five levels defined by this item departed significantly from linearity. Intervals between adjacent response categories were unequal (Davies & Ware, 1981). Accordingly, these response options for the General Health Perceptions Scale were recalibrated for the RAND–36 HSI.

Simple summation scoring was also questioned for several of the SF–36 items by Hays et al. (1993). Haley, McHorney, and Ware (1994) employed the Rasch model to examine the hierarchical structure, unidimensionality, and reproducibility of item positions (calibrations) on

the 10-item Physical Functioning Scale. This analysis generated an empirical item hierarchy, confirmed the unidimensionality of the scale for most respondents, and established the reproducibility of item calibrations across patient populations and repeated tests. More recently, McHorney, Haley, and Ware (1997) compared simple summation scoring based on the Likert scale with the Rasch IRT scaling model for the 10-item Physical Functioning Scale. Findings favored the Rasch model in discriminating between patients who differed in disease severity. Differences were reported as most apparent in clinical groups whose scores approximated the extremes of the distribution. It was suggested that the Rasch model of scoring would be relevant to the clinical interpretation of individual scores on this scale. The development of the RAND–36 HSI stems from the significant amount of early work documenting the need for revision in scoring.

Application of IRT Weighting

IRT methodology was used for scoring the RAND–36 HSI responses in order to take into account relative item weights within each scale and item response weights within each item simultaneously.

The first step in applying the IRT scoring method was to select an appropriate sample, that is, one for which there were no missing responses to any items. Of the age-based standardization sample, 737 protocols met this criterion. Based on this sample's responses, the IRT weighting for items in each of the eight RAND–36 HSI scales was determined by the following methodology.

Calibration based on a one-parameter IRT logistic model was conducted to obtain the item characteristic curve (ICC) for each item in the scale. An ICC provides the expected item score (item response option) as a function of the individual's ability level on the construct being measured (the latent-trait continuum), given the item difficulty (or step values) of that item. For each ICC, the corresponding ability level for each response option can be obtained. Once all response options were placed on the same latent-trait continuum, they were rescaled to a 0–100 linear scale where the existing minimum ability level was set to 0 and the maximum ability level was set to 100. This new scale served as the basis for the IRT weighting of item responses.

A raw score is computed by summing the IRT weight for the selected response option for each item on the scale. Appendix A of this Manual presents the procedures and tables for these computations; Appendix B presents the procedures and tables for computing scale and composite *T* scores (the derivation of *T* scores is described in Chapter 4).

Effects of IRT Scoring on Distribution

Initial expectations were that the differences in the distribution due to IRT scoring would not be equal for all scales. Moreover, differences would be greatest for those scales with the widest range of ability levels, as determined by the number of items and the number of response options for items on that scale. Therefore, scales composed of items with only two response options, such as the 4-item Role Limitations due to Physical Health Problems and the 3-item Role Limitations due to Emotional Problems, were anticipated to show little

change in score distribution due to IRT scoring. Methods of identifying potential differences between simple summation and IRT scoring of scales included comparison of means, skewness, and kurtosis, as well as chi-square analysis.

For purposes of comparison, IRT raw scale scores were placed on the same 0–100 range as the scores derived by the simple summation method. A comparison of raw-score means based on the IRT and simple summation methods yielded significant differences for six of the eight scales. As predicted, those scales offering two response options per item did not show a significant difference (i.e., Role Limitations due to Physical Health Problems and Role Limitations due to Emotional Problems). On all six of the other scales, the IRT method yielded mean scores significantly lower than the mean scores derived with the simple summation method ($p < .0001$). Among the scales of the Physical Health Composite, Physical Functioning showed a small but significant difference between mean scores, and on Pain and General Health Perceptions, mean scores obtained by the IRT method were more than 3 points lower than mean scores obtained by simple summation. Among the Mental Health Composite scales, mean scores on Emotional Well-Being and Energy/Fatigue obtained by the IRT method were 9 and 7 points lower, respectively, than mean scores based on simple summation.

Some differences in distribution of scores were also indicated. A comparison of skewness and kurtosis demonstrated that IRT scoring yields smaller skewness (absolute value) and smaller kurtosis for seven of the eight scales. Skewness was consistently smaller with IRT scoring; the largest differences were for Pain (0.23), General Health Perceptions (0.20), and Energy/Fatigue (0.43). Kurtosis was also consistently smaller with IRT scoring. These differences were statistically significant ($p < .01$) for Physical Functioning (0.47), Pain (0.70), General Health Perceptions (0.45), Emotional Well-Being (1.58), and Energy/Fatigue (0.68). These results indicate that IRT scoring generally resulted in less skewness and smaller kurtosis, that is, in distributions that were more spread out and flatter.

The precise nature of the shift in distribution afforded by IRT scoring varied across scales. Chi-square analyses of differences by scoring method (IRT or simple summation) across scales, with scores organized into 10 ability levels, revealed shifts in the distribution of scores that were significant in five of eight scales: Pain ($\chi^2 = 46.69$, $p < .001$), General Health Perceptions ($\chi^2 = 17.08$, $p < .05$), Emotional Well-Being ($\chi^2 = 104.53$, $p < .001$), Social Functioning ($\chi^2 = 146.49$, $p < .001$), and Energy/Fatigue ($\chi^2 = 39.17$, $p < .001$).

Development of the Composites

The theoretical assumptions underlying the composite scores, as well as their psychometric development are discussed here. The steps for computing these composite scores are presented in Appendix B.

Studies of health status have consistently identified distinct physical health and mental health factors. These factors have been identified in patient groups (Hays, Marshall, et al., 1994; McHorney et al., 1993; Ware, Gandek, & the IQOLA Project Group, 1994), in the general U.S. population (Ware et al., 1993), and across different demographic and patient groups (Ware, Kosinski, Bayliss, et al., 1995).

The methodology used for deriving composite scores for the RAND–36 HSI differs from that employed with the SF–36 in several ways (Ware, Kosinski, & Keller, 1994). First, the method used for the RAND–36 HSI is based on results from principal axis factor analysis rather than principal components factor analysis, which was applied to the SF–36 (Ware, Kosinski, et al., 1994). With the principal axis factor analysis, the obtained factors (physical health and mental health) are based on common, rather than on total variance among scales (Gorsuch, 1983); they are the true underlying factors (as opposed to sample-specific components); and they explain as much of the common variance as possible.

Second, the method of composite score construction used for the RAND–36 HSI differs from those presented previously (Ware, Kosinski, et al., 1994) because the formula for the composite score includes only those scales that load highly on that factor. As a result, the Physical Health Composite score is derived from scores on the Physical Functioning, Role Limitations due to Physical Health Problems, Pain, and General Health Perceptions scales. The Mental Health Composite score is derived from the scores on the Emotional Well-Being, Role Limitations due to Emotional Problems, Social Functioning, and Energy/Fatigue scales.

An additional difference in the methodology used for the RAND–36 HSI is that it employs an oblique rotation rather than the orthogonal rotation employed previously (Ware, Kosinski, et al., 1994). The oblique rotation method is based on the assumption that the physical health and mental health factors are correlated, not independent, as assumed in orthogonal rotation methods. Previous research has found that physical and mental aspects of health are distinguishable, but also significantly correlated (Hays, Marshall, et al., 1994).

The existence of distinct physical and mental components of health status has been well documented (Hays, Marshall, et al., 1994; McHorney et al., 1993). In addition, factor patterns found in previous analyses have replicated the significant loadings on physical health and mental health factors. In previous studies that included all eight scales in the calculation of each composite, three scales—General Health Perceptions, Social Functioning, and Energy/Fatigue—loaded on both factors. This result is cited as justification for their inclusion in both the Physical Health and Mental Health composites. The composite scores for the RAND–36 HSI do not include overlapping scales because the factor loadings did not warrant their inclusion and because the Global Health Composite was developed to represent the overlapping aspects of physical health and mental health.

Thus, in addition to the Physical Health and Mental Health composites, the RAND–36 HSI yields a Global Health Composite. This composite reflects the conception of underlying global health that is composed of both physical health and mental health and potentially overlapping aspects. It can be viewed as a "thermometer" of general health. This composite is consistent with the original conception of general health as an integrative, underlying construct. In practice, the Global Health Composite score may be used in circumstances when one measure of general health status is required or when the distinction between physical health and mental health is not important.

Physical Health and Mental Health composites were derived by common factor analysis (principal axis method with two iterations and squared multiple correlations for priors) that specified two related factors with oblique rotation. For the analysis, scores on all eight scales

obtained by the age-stratified sample (N = 500) were used. The analysis was restricted to two iterations because when factors are allowed to be correlated, the model, which will continue to change through iterations, may be overfitted and lead to estimates of communalities greater than 1 (actual communalities cannot exceed 1). Such communalities are known as "Heywood" cases. Limiting the number of iterations to two with good prior estimates leads to more accurate estimates of the communalities (Gorsuch, 1983). The rotated promax factor pattern matrix presented in Table 3.1 reveals that the four scales related to physical health defined the first factor, with loadings ranging from .63 to .90. This factor was defined as the Physical Health Composite. The four scales related to mental health loaded on the second factor, with loadings ranging from .53 to .95. This factor was defined as the Mental Health Composite.

Factor scores for both the Physical Health Composite and the Mental Health Composite were derived for each member of the age-stratified sample with the UniMult program (Gorsuch, 1991). The third composite, the Global Health Composite, was then derived by factoring the two factor scores, physical health and mental health, with one common factor specified.

Table 3.1. Promax Factor Pattern Loadings for RAND–36 HSI Scales

Scale	Factor 1 Physical Health	Factor 2 Mental Health
Physical Functioning	**.90**	−.21
Role Limitations due to Physical Health Problems	**.79**	.00
Pain	**.76**	.02
General Health Perceptions	**.63**	.18
Emotional Well-Being	−.21	**.95**
Role Limitations due to Emotional Problems	.00	**.59**
Social Functioning	.30	**.53**
Energy/Fatigue	.24	**.58**

Note. N = 500. The factor analysis was based on the scores obtained by the age-stratified sample. The estimated correlation between Factors 1 and 2 was .66.

A linear equation containing beta weights was obtained for each composite by regression analysis. For the analysis, the factor score of the composite was used as the dependent variable, and the scores of the scales contributing to that composite were used as the independent variables. It should be noted that the scale weights developed from the regression equations do not match the rank ordering of the factor loadings because scale weights represent the relative contribution of what is measured by each scale to the factor, whereas the factor loadings represent the contribution of the underlying factor to what is measured by each scale.