Validation of an Interval Scaling: The Sickness Impact Profile

by William B. Carter, Ruth A. Bobbitt, Marilyn Bergner, and Betty S. Gilson

The Sickness Impact Profile (SIP) is a measure of sickness-related behavioral dysfunction consisting of 189 items in 14 topic categories. To increase its discrimination, precision, and sensitivity in accounting for variance, the decision was made to scale the instrument. A two-step direct scaling procedure was used in order to avoid the monumental scaling tasks required by indirect procedures that guarantee equal-interval results; but because an equal-interval scale was needed, it was necessary to validate the scale values obtained and investigate the equal-interval properties of the obtained scale. A three-stage validation process is described, consisting of an initial scaling by a group of 25 health professionals and students in 1973, a second scaling by 108 members of a prepaid group health plan in 1975, and an investigation of the metric properties of the resulting scale values. In addition, the concept of dysfunction underlying the SIP was validated. SIP scores from a field trial were compared with mean ratings of severity of dysfunction represented by the combinations of checked items from which the scores were derived.

The purposes of this article are to elaborate on some of the conceptual and methodological issues of scaling the Sickness Impact Profile (SIP), to describe two validations of a preliminary scaling of the SIP and a validation of the metric obtained from the scaling procedure, and to discuss results of these validations in relation to the conceptual issues.

Background

The SIP was conceptualized as a behaviorally based measure of sickness-related dysfunction that would provide an appropriate and sensitive measure of health status designed to aid in assessing the outcomes of health care services [1]. Work on the SIP began in 1972 with the development of procedures to

Research supported by grant number HS-01769 from the National Center for Health Services Research, Health Resources Administration.

Address communications and requests for reprints to William B. Carter, Ph.D., Department of Health Services, School of Public Health and Community Medicine, University of Washington, Seattle, WA 98195. Ruth A. Bobbitt, Ph.D., Marilyn Bergner, Ph.D., and Betty S. Gilson, M.D., are members of the department of health services at the University of Washington.

Table 1. The 14 Behavior Categories in the Sickness Impact Profile, with Sample Items from Each Category

Social interaction	I make many demands, e.g., insist that people do things for me, tell them how to do things
	I am going out less to visit people
Ambulation	I am walking shorter distances
	I do not walk at all
Sleep and rest	I lie down to rest more often during the day
	I sit around half asleep
Taking nutrition	I am eating no food at all, nutrition is taken through tubes or intravenous fluids
	I am eating special or different food, e.g., soft food, bland diet, low salt, low fat foods
Usual daily work	I often act irritable toward my work associates, e.g., snap at them, give sharp answers, criticize easily
	I am not working at all
Household management	I have given up taking care of personal or household business affairs, e.g., paying bills, banking, working on budget
	I am doing less of the regular daily work around the house than I usually do
Mobility/confinement	I stay in one room
	I stop often when traveling because of health problems
Movement of the body	I am in a restricted position all the time
	I sit down, lie down, or get up only with help
Communication	I communicate only by gestures, e.g., moving head, pointing, sign language
	I often lose control of my voice when I talk, e.g., my voice gets louder, starts trembling, changes pitch
Pastimes, recreation	I am doing more physically inactive pastimes instead of my other usual activities
	I am going out for entertainment less often
Intellectual function	I have difficulty reasoning and solving problems, e.g., making plans, making decisions, learning new things
	I sometimes behave as if I were confused or disoriented in place or time, e.g., where I am, who is around, directions, what day it is
Family interaction	I isolate myself as much as I can from rest of family
	I am not doing the things I usually do to take care of my children or family
Emotions, sensations	I act irritable and impatient with myself, e.g., talk badly about myself, swear at myself, blame myself for things that happen
	I laugh and cry suddenly for no reason
Personal hygiene	I dress myself, but do so very slowly
	I do not have control of my bowels

collect and evaluate statements describing behavioral dysfunction obtained from patients, individuals caring for patients, the apparently healthy, and health care professionals [2]. The results of these procedures yielded 312 unique statements or items which were subsequently subjected to a standard sorting process. Fourteen categories resulted, each appearing to describe a specific type of activity. A summary of these categories with sample items from each is shown in Table 1. The items were included in a structured interview format

in which subjects were asked to respond only to those items that they were sure described them and were related to their health.

Rationale for Scaling the SIP

The decision to scale SIP items was based on consideration of the content of the SIP and the desired use of the instrument as a measuring tool. First, the SIP contains statements that cover a broad range of subject matter and a diversity of severity within each subject area. For example, the ambulation category includes the following two items: "I am walking shorter distances" and "I do not walk at all." These two statements describe two quite different levels of behavioral dysfunction. A similar range occurs among statements in other categories concerning different content areas. Thus one finds intuitive evidence suggesting that SIP items should not be equally weighted.

A second reason for scaling the SIP was the limited discriminative capacity inherent in the number of responses from different subjects. Because of the broad range of behavioral dysfunctions included in the SIP, the relative impact of sickness on the behavior of two persons could be quite different even though they both checked the same number of descriptive items [3]. For example, it is quite possible that an individual suffering minor dysfunctions expressed as slight slowing down in usual activities might check as many items as an individual who indicates that certain usual functions are not being performed at all. This lack of sensitivity and discriminative ability may be a function of the lack of weights for the various behavioral expressions of the relative severity of dysfunction described by various items.

A third consideration for scaling is the possibility of social or cultural differences in the perception of health-related behaviors. That is, individual items may be valued differently by different subgroups of the population. Scaling provides an opportunity to test the extent to which a single instrument is valid across social and cultural groups. If differences should be observed, then application of the appropriate set of scale values would maintain the instrument's sensitivity.

A fourth reason for scaling derives from the criticisms that have been made of many previous attempts at evaluating outcomes of health care. Among many such criticisms [2,4], one commonly cited is the insensitivity of the measures used. Scaling is designed explicitly to increase a measuring instrument's precision and sensitivity in accounting for variance [5]. Although a straight count of items checked may account for the bulk of the variance, the crucial level of precision required in making fine discriminations may be attainable only through the refinement of the instrument provided by scaling. Thus, even though the utility of scaling cannot be determined beforehand, in our opinion it is a necessary procedure that merits the time and costs involved.

Selecting a Scaling Method

Many possible scaling methods will provide at least an equal-interval metric, given that certain assumptions are satisfied. Engen [6] has classified these

methods into direct and indirect scaling procedures. In indirect methods judges are required to make only ordinal ratings, and the metric scale values are derived by the researcher from the ordinal judgments. These methods require that the researcher be willing to make stronger measurement assumptions about the conceptual continuum than about a judge's rating ability. Unfortunately, indirect methods are practicable only in situations where a small number of stimuli are to be rated. For example, in the method of paired comparisons, a frequently used indirect scaling procedure, judges are required to compare each stimulus with every other. Thus, for a 200-item instrument, 19,900 individual comparisons would be required.

Direct scaling methods, on the other hand, require the judges to make metric ratings, which simplifies the step between the raw data and the final scale values. These methods are easily applied to a large number of stimuli or items and require the researcher to make only one assumption, namely, that the judge is able to make his ratings at the quantitative level requested by the instructions. For the purpose of scaling the SIP the method of equal-appearing intervals was selected as the most tractable direct scaling procedure for providing a metric scale.

Preliminary Scaling of the SIP, 1973

Sample and Procedure

As part of the initial development of the SIP, items were scaled in 1973 by a group of 25 judges including physicians, nurses, and health administration students. To avoid overwhelming judges with the task of making discriminations among 312 items across all categories, a two-step scaling procedure was employed. In the first step, judges rated items within each SIP category. The judging sessions were standardized. Categories and items within each category were shuffled prior to each presentation. Instructions were read to each judge individually at the beginning of his/her session, and a copy of the instructions was left for the judge's reference during the session.

Each judge was instructed to rate, on an equal-interval 11-point scale, the extent to which each questionnaire item within a given category described a dysfunction in behavior. The scale ranged from "minimally dysfunctional" to "severely dysfunctional," and dysfunction was specifically defined. The properties of the equal-interval scale were stressed, and an example was given to illustrate its use. Furthermore, judges were asked to rate the severity of the dysfunction described in an item without regard for what might be causing it.

After all of the items within a given category had been placed along the 11-point scale, judges were asked to review the items they had placed at each scale point to ensure that those items were more similar to each other in terms of degree of dysfunction than they were to items placed at any other scale point. Judges were encouraged to correct any discrepancies they observed and, after they were satisfied with the arrangement, to record their judgments on the response sheet provided. After this recording was completed, each judge

received the next category of items and the same procedure was followed. Judges were asked to rate the items in each category without regard for the way they had distributed items in the other categories. This procedure continued until each judge completed 14 categories.

Then, in order to permit comparison of items within and across categories, the same judges were asked to perform a second rating task. Using the same scaling procedures, they rated on a single 15-point scale the two items from each category that had previously been rated by the group as least and most dysfunctional.

The Reliability of Scaling

An obtained scale value is valid only to the extent to which an item is sufficiently definitive and clearly stated that a representative group of judges can agree on its relative scale value. The primary focus in scaling the SIP was to secure a consensus about the scale values assigned to items. This approach has been referred to in the literature as stimulus-centered, as opposed to subject-centered [7,8]. Such a stimulus-centered focus necessitated an assessment of reliability and consistency among judges and a determination of possible systematic variations within and across judgment groups. Reliability in this context is therefore determined in terms of group consensus. We computed a correlation coefficient comparing each judge's ratings with the group mean ratings across items, which provided an index reflecting the degree to which the pattern of each judge's ratings corresponded to those of the group. All the 1973 judges were found to be reliable, in that none assigned ratings that were widely disparate from the mean ratings of the group.

The next step was to identify any items in the SIP that were not clearly stated or definitive enough that this group of judges could reach a consensus about them. A standard normal deviate score was computed for each item by treating each item's standard deviation as one would treat any ordinary score, that is

$$Z_i = \frac{SD_i - \overline{SD}}{S_{SD}}$$

where Z_i = the standard normal deviate for item i
SD_i = the standard deviation of item i
$\overline{SD}$ = the mean of all the item standard deviations
S_{SD} = the standard deviation of all item standard deviations

The size and sign of this deviate score provide a relative index of rating difficulty for a given item. That is, if an item were difficult to rate, one would expect little agreement among judges regarding the scale values assigned that item, which would be reflected as a large positive deviate score. For example, a deviate score of +1.5 for an item means that that item had a standard deviation one and one-half times larger than the average on the other items in the instrument. Following this logic, a large negative deviate score is indicative of a

strong degree of consensus; a deviate score of –1.5 means that the item had a standard deviation one and one-half times smaller than the average of the other items in the instrument. No firm statistical criteria are available to decide what degree of difficulty is acceptable. This decision was based on global criteria inherent in the instrument being developed.

In this analysis 284 items were found to be reliable; the remaining 28 were either dropped or revised and subsequently rescaled.

Preliminary Validation of the 1973 Scale Values

At this stage in the development of the SIP, data were also collected to test scoring based on these scale values and to validate the construct of dysfunction. A separate group of judges rated SIP profiles obtained from field trial subjects in terms of dysfunction. The ratings of the profiles and the scores assigned to them on the basis of the scale values of items checked were highly correlated. Analysis of the judges' ratings indicated that both the number of items checked and the severity of dysfunction expressed by those items were dominant factors and contributed to the judgments made of the SIPs [1].

This initial validation represented only the first stage of a proposed three-stage validation process. The second stage involved a validation of the preliminary (1973) scale values by replicating the scaling procedures, this time with a group of judges drawn from a health care consumer population. The third stage involved a validation of the metric obtained from the equal-interval scaling procedure. The second and third stages of the validation were completed in 1975.

Consumer Scaling of the SIP, 1975

Sample and Procedure

A random sample of 173 persons was drawn in January 1975 from the enrollees of Group Health Cooperative (GHC), a prepaid group practice in Seattle, Washington. The sample was stratified by age (18–44, 45–64, 65–74), sex, and type of membership. (GHC has two basic types of enrollees: co-op members, who pay an initial entrance fee, have voting privileges, and have the possibility of lifetime coverage, and nonco-op members, who join through a group contract at their place of employment.) Persons in each of the 12 resulting strata were randomly contacted until nine persons in each stratum agreed to participate in the study (N = 108). Each participant was reimbursed $30 for completing two ratings tasks that required a total of three hours. Response rates were lowest for males over age 44 and for females between the ages of 18 and 44. One of the judges failed to follow instructions and was subsequently replaced.

The version of the SIP used—the current version—consists of 189 items grouped into 14 categories or areas of living. The scaling procedures employed replicated those used in obtaining the 1973 scale values; however, the judging task was simplified and shortened by dividing the 14 item categories into two

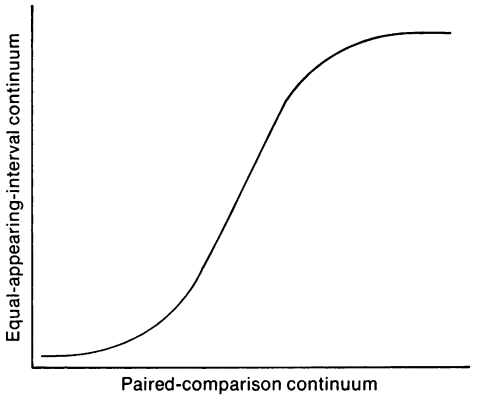

Fig. 1. Theoretical relationship between scale values derived from the method of paired comparisons and the method of equal-appearing intervals for the same set of items [9].

content-similar clusters. Persons in each stratum of the consumer sample were randomly assigned to one of two judgment groups.

The reliability of the 1975 scale values was analyzed by the same methods used in 1973. The 1973 and 1975 scale values were compared by regression analysis.

Validating the Equal-interval Scaling Metric

In the use of a direct scaling technique it was assumed that judges would be able to make quantitative ratings at the level requested by the instructions. This is not a trivial assumption since most statistical analyses pivot on this very factor. Furthermore, it is not uncommon to find ratings of this type that deviate from equal intervals. For example, Edwards [9] summarized data originally reported by Hevner [10], who found that when the results of an equal-appearing interval scaling procedure were plotted against the scale values assigned to the same stimuli by the method of paired comparisons (which guarantees an equal-interval scale), the relationship between the two sets of scale values was not linear at the two extremes of the equal-appearing-interval continuum. This relationship is shown in Fig. 1. The departure from linearity invalidates the metric properties along the extremes of the rating scale.

To avoid these problems, Edwards [9] favored the use of Thurstone's method of successive intervals, which transforms equal-appearing-interval ratings into successive interval values. This method was chosen to test the SIP scaling, since it offered the dual option of either validating the scaling procedure used, that is, validating the assumption that judges were able to follow the equal-interval instructions, or, if this assumption were in error, of transforming the obtained scale values into successive interval values [11].

Table 2. Quartile Summary of the Range of Correlations of Each Judge's Ratings with the Group Mean Ratings Across Items

Quartile	Range of correlations	Number of judges
1 (lowest)	–0.410–0.210	2
	–0.200–0.000	2
	0.010–0.090	1
	0.260–0.580	22
2	0.590–0.715	27
3	0.716–0.787	27
4 (highest)	0.788–0.878	27

Thus we reasoned that since the method of successive intervals yields a derived equal-interval scale, then a regression analysis of those derived values and the mean scale values assigned by the judges should result in a linear relationship if the judges' mean values gave a metric scale. (Mean and median scale values were found to be equivalent, so mean values were used for analysis because they were computationally convenient.) However, if the judges' average ratings departed from an equal-interval metric, then the regression would yield a relationship similar to that shown in Fig. 1.

Results

Consumer Judge Reliability and Item Rating Difficulty

Setting a criterion for judge reliability is difficult, particularly when one wants to fairly represent the judgment group. The co-criteria of consistency and fair representation seem to be inherently incongruent. If consistency were the primary focus, an a priori statistical criterion could be established; if fair representation were the primary concern, there would be no need to assess reliability. Taken together, however, these two concerns indicate that those judges with scores that are widely disparate from the group mean can be dropped as unreliable, and the criteria are established post hoc.

A quartile breakdown of the correlations between judges' ratings and the group mean is shown in Table 2. In the first (lowest) quartile, five of the judges had correlations ranging from –0.41 to 0.09. Six of the remaining 22 judges in the first quartile had correlations from 0.26 to 0.36. After a careful review of the correlations, the judges' individual response patterns, and comments from the 11 lowest judges, it became clear that the five judges with the lowest correlations had not understood the scaling task. Therefore these five judges were dropped from further analysis.

The mean ratings of judges in each stratum were compared using analysis of variance. No significant differences were found. However, because of the small sample size within each stratum, group means were collapsed across

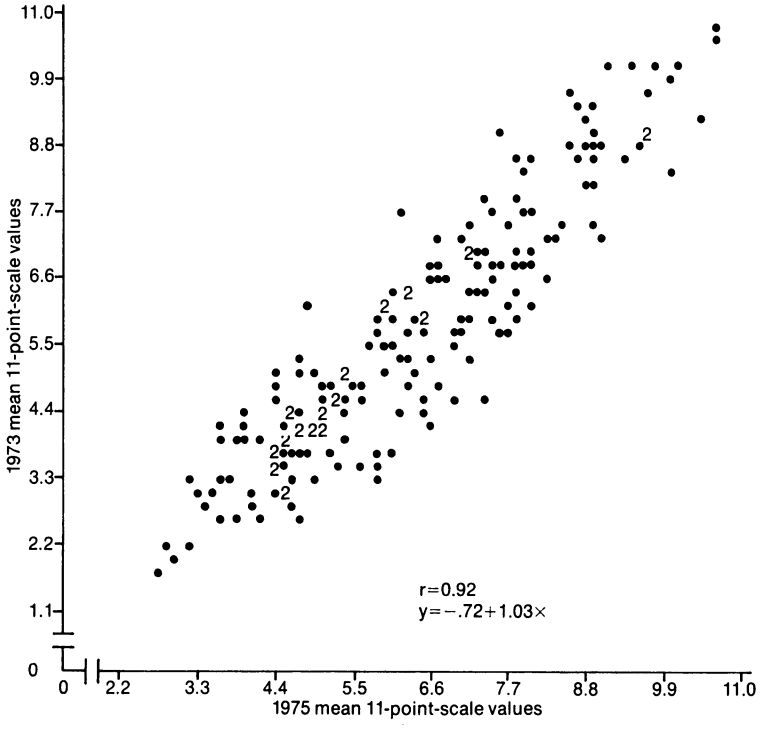

Fig. 2. Comparison of the mean 11-point-scale values assigned to 189 SIP items by the 1973 and the 1975 judges. Number beside a point shows number of observations at that point.

membership type, age, and sex in separate computations. Again, no differences were found.

Analysis of item Z-scores indicated that 23 items were difficult to rate. These items are presently being revised, and at some point in 1977 they will be re-scaled using a similar procedure.

Comparison of 1973 and 1975 Scale Values

Validation of the 1973 scale values was based on regression analysis. Item scale values from the 1973 and 1975 scalings are plotted in Fig. 2; the correlation between the two sets of values is 0.92 ($p \leqslant 0.00001$), with a slope of 1.03 ($p \leqslant 0.00001$) and an intercept of –0.72 ($p \leqslant 0.0004$). These results show a striking similarity between two quite different judgment groups. The only real divergence between the groups was that the consumer judges tended to assign higher scale values. The reasons for such a difference can only be conjectured at present. The important point, however, is that this inflational factor was essentially constant throughout the scale: the relationship between items did not change.

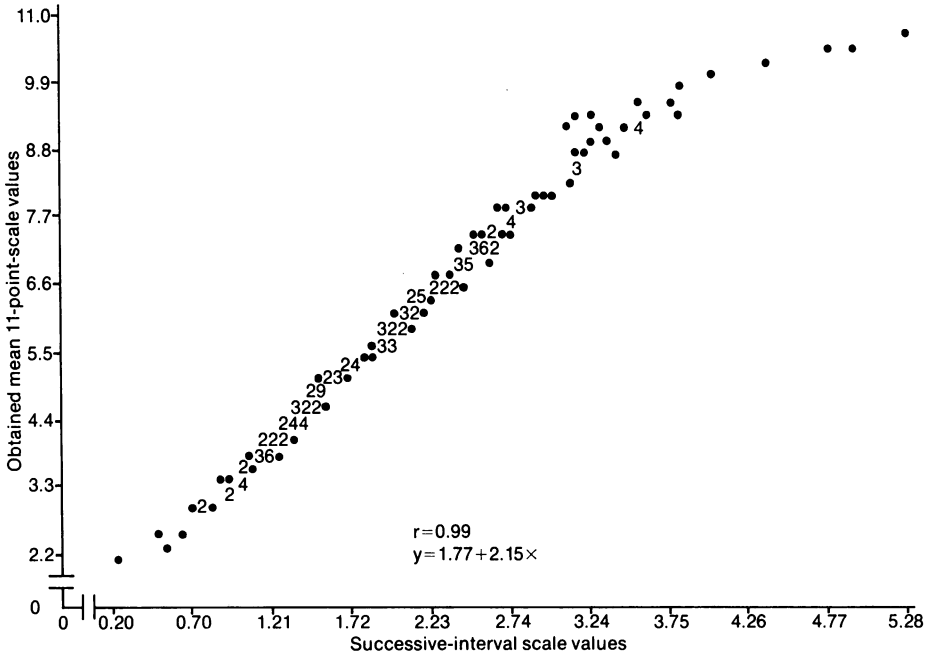

Fig. 3. Comparison of the obtained mean 11-point-scale values from the combined group of judges and the scale values derived from the method of successive intervals for the 189 SIP items. Number beside a point shows number of observations at that point.

On the basis of these results and the desire to have the SIP scale values represent as diverse a group as possible, a decision was made to combine the 1973 and 1975 judges' ratings to compute a final set of SIP item scale values.

Analysis of the Equal-interval Metric

Thurstone's method of successive intervals was applied to 284 reliably scaled items from the 1973 scaling for a preliminary test of the scaling metric. The results indicated that the values derived by Thurstone's method were comparable throughout the 11-point scale to those assigned by the judges using the equal-interval category scaling method, thus validating the 1973 scaling methodology. It was assumed that this scaling methodology was also valid in 1975. The 1973 and 1975 judges' ratings were combined for the 189 items common to both scaling studies, and the results were regressed against the successive-interval values computed by Thurstone's method. The regression line is presented in Fig. 3; the digits associated with a point show the number of observations represented by that point.

For this particular equal-interval assessment, the correlation coefficient is not the validating criterion. Rather, the slope of the regression line and the

shape of the plotted function provide the equal-interval criteria. The actual value of the slope is partially dependent on the numerical range of the scale values being evaluated. For example, a perfect linear relationship between two sets of values with the same numerical range would result in a slope of 1.0. Given the 11-point range of obtained values and the computed range of successive interval values presented in Fig. 3, a perfect linear relationship would yield a slope of 2.17. The slope revealed by the data presented in Fig. 3 is 2.15. In addition, with the exception of the five most extreme items, no significant departure from a linear trend was observed. Thus both the slope and shape provide substantial evidence that the judges by and large employed an equal-interval scale in making their judgments of dysfunction.

Discussion

Many investigators have been unable to obtain metric scales from direct scaling procedures and have attributed this to the judges' inability to discriminate with the required precision. However, Edwards [9] has postulated that a relationship like that shown in Fig. 1 may result from an inadequate scaling procedure. For example, it is quite possible for a judge to assign an extreme item the maximum scale value early in the scaling task and later encounter an item more extreme than the first. Since there is no more extreme scale point, the judge may assign the same maximum value to both items. In other words, the problem is not one of discriminative ability, but rather that the judge is not forced to compare these two items as would be the case in the method of paired comparisons. Data reported by Hevner [10] and Kelley, Hovland, Schwartz, and Abelson [12] show that this indeed happens.

The scaling methodology employed was specifically designed to avoid many of the pitfalls previously elaborated in the scaling literature. For example: the judges were asked to compare the items placed at each scale point to ensure that those items were more similar to each other in terms of degree of dysfunction than they were to items placed at any other scale point; the scale was of sufficient size to allow for an appropriate level of discrimination; the scale endpoints were defined to provide judges a meaningful referent; the number of items being rated at any one time was minimized; and unreliable judges were excluded from these analyses. All of these factors undoubtedly had some impact on the end result.

As can be seen in Fig. 3, however, a few of the most extreme items did deviate from linearity, perhaps because of the nature of the items themselves. That is, the most extreme SIP item—"I do not move any part of my body"—may be so extreme as to extend beyond the limits that bind the other items in the instrument.

In developing the scaling methodology used for the SIP there was particular concern about the metric obtained for such items. Thus, in order to take account of extreme items and to be able to compare items within and across categories, a second scaling task was undertaken [1]. In the first task, judges

rated items separately for each category. From these judgments, item means were computed and the average least and most dysfunctional items were selected as stimuli for the second task. In the second scaling task, the same judges were asked to rate these least and most dysfunctional items on a single 15-point scale. The resulting endpoint scale values were used to derive 15-point scale values for all items.

The results of the second rating task, which have not been reported here, provided some evidence to support the hypothesis that the extreme items mentioned did extend beyond the 11-point scale. The five most severely rated items shown in Fig. 3 received the highest mean ratings on the 15-point scale. However, the degree of separation among these five items is still insufficient to provide a perfect linear fit, and the disposition of these items is presently being considered. In any event, the striking degree of linearity over the remaining portions of the scale provides sufficient evidence to validate the scaling metric.

The primary purpose for scaling the SIP was to increase its precision and sensitivity in accounting for variance. Data from the 1974 field trial have shown the scored SIP to be a useful tool in discriminating among groups that varied in type and severity of dysfunction [13]. In the current 1975–76 field trial, a number of diagnostic groups are being followed throughout a course of clinical treatment; the results will provide additional tests of sensitivity.

In general, the utility of a scaling procedure transcends its primary goal by providing the researcher with a great deal of data about the instrument being scaled. Although time constraints have not permitted the discussion of many types of information gleaned from our scaling studies, some of the data presented allowed the identification of items that were particularly difficult to judge. This information has proved to be very useful in item revisions, and the 1975 consumer data have been incorporated into the current revision process.

Rescaling the SIP has allowed us to validate the scale values obtained in the preliminary item scaling. The consistency of the relative scale values assigned to items by two different groups of judges is particularly striking. These data provide evidence of the reliability of the procedures used and indicate that the instrument and procedure are relatively stable to the cultural biases of these two groups.

As part of the future research planned in the development of the SIP, a Chicano-Spanish version has been translated and will be scaled by a consumer group. A comparison of these scale values with the ones previously obtained will provide further opportunity to test the extent to which the SIP and the scale values are valid across social and cultural groups.

REFERENCES

1. Bergner, M., R.A. Bobbitt, S. Kressel, W.E. Pollard, B.S. Gilson, and J. Morris. The Sickness Impact Profile: Conceptual formulation and methodological development of a health status index. *Int J Health Serv* 6:393 Summer 1976.
2. Gilson, B.S., J.S. Gilson, M. Bergner, R.A. Bobbitt, S. Kressel, W.E. Pollard, and M. Vesselago. The Sickness Impact Profile: Development of an outcome measure of health care. *Am J Public Health* 65:1304 Dec. 1975.

3. Gilson, B.S., M. Bergner, R.A. Bobbitt, W.E. Pollard, and D. Martin. Revision and test of the Sickness Impact Profile, 1973–74. Department of Health Services, School of Public Health and Community Medicine, University of Washington, Seattle, 1974.
4. Elinson, J. Effectiveness of Social Action Programs in Health and Welfare. In A. B. Bergman (ed.), *Report of the Fifty-sixth Ross Conference on Pediatric Research: Assessing the Effectiveness of Child Health Services*, pp. 77–88. Columbus, OH: Ross Laboratories, Inc., 1967.
5. Maranell, G.M. (ed.). *Scaling: A Sourcebook for Behavioral Scientists.* Chicago: Aldine, 1974.
6. Engen, T. Psychophysics II. Scaling Methods. In J.W. Kling and L.A. Riggs (eds.), *Woodworth and Schlosberg's Experimental Psychology, Volume I: Sensation and Perception*, pp. 47–86. New York: Holt, Rinehart and Winston, 1972.
7. Torgerson, W.S. *Theory and Methods of Scaling*, pp. 45–48. New York: Wiley, 1958.
8. Shinn, A.M. Jr. Relations Between Scales. In H.M. Blalock Jr. (ed.), *Measurement in the Social Sciences: Theory and Strategies*, pp. 121–158. Chicago: Aldine, 1974.
9. Edwards, A.L. *Techniques of Attitude Scale Construction*, pp. 120–147. New York: Appleton-Century-Crofts, 1957.
10. Hevner, K. An empirical study of three psychophysical methods. *J Gen Psychol* 4:191 Dec. 1930.
11. Blischke, W.R., J.W. Bush, and R.M. Kaplan. Successive intervals analysis of preference measures in a health status index. *Health Serv Res* 10:181 Summer 1975.
12. Kelley, H.H., C.I. Hovland, M. Schwartz, and R.P. Abelson. The influence of judges' attitudes in three methods of scaling. *J Soc Psychol* 42:147 Aug. 1955.
13. Bergner, M., R.A. Bobbitt, W.E. Pollard, D.P. Martin, and B.S. Gilson. The Sickness Impact Profile: Validation of a health status measure. *Med Care* 14:57 Jan. 1976.