The Sickness Impact Profile: Development and Final Revision of a Health Status Measure
Author(s): Marilyn Bergner, Ruth A. Bobbitt, William B. Carter and Betty S. Gilson
Source: *Medical Care*, Vol. 19, No. 8 (Aug., 1981), pp. 787-805
Published by: Lippincott Williams & Wilkins
Stable URL: https://www.jstor.org/stable/3764241
Accessed: 25-01-2019 10:15 UTC

MEDICAL CARE
August 1981, Vol. XIX, No. 8

Original Articles

The Sickness Impact Profile: Development and Final Revision of a Health Status Measure

MARILYN BERGNER, PH.D., AND RUTH A. BOBBITT, PH.D., WITH WILLIAM B. CARTER, PH.D., AND BETTY S. GILSON, M.D.

The final development of the Sickness Impact Profile (SIP), a behaviorally based measure of health status, is presented. A large field trial on a random sample of prepaid group practice enrollees and smaller trials on samples of patients with hyperthyroidism, rheumatoid arthritis and hip replacements were undertaken to assess reliability and validity of the SIP and provide data for category and item analyses. Test–retest reliability (r = 0.92) and internal consistency (r = 0.94) were high. Convergent and discriminant validity was evaluated using the multitrait–multimethod technique. Clinical validity was assessed by determining the relationship between clinical measures of disease and the SIP scores. The relationship between the SIP and criterion measures were moderate to high and in the direction hypotheszed. A technique for describing and assessing similarities and differences among groups was developed using profile and pattern analysis. The final SIP contains 136 items in 12 categories. Overall, category, and dimension scores may be calculated.

THIS ARTICLE provides an overview of a 6-year research project undertaken to develop a behaviorally based measure of health status, the Sickness Impact Profile (SIP). Since preliminary and interim accounts of the methodology, testing and development of this measure have been reported previously, this article will summarize the early work and emphasize the final phases of development.

Purpose of the SIP

The SIP was developed to provide a measure of perceived health status that is sensitive enough to detect changes or differences in health status that occur over time or between groups. It was designed to be broadly applicable across types and severities of illness and across demographic and cultural subgroups. The SIP is intended to provide a measure of the effects or outcomes of health care that can be used for evaluation, program planning and policy formulation. Since sensitivity to minimal levels of dysfunction is critical to reliable and valid estimates of historical

From the Department of Health Services, University of Washington, Seattle.

This research was supported by the U.S. Department of Health, Education, and Welfare, Public Health Service, National Center for Health Services Research, under grant no. HS 01769 and contract no. 106 HSR 283(4), and by the HMO Service of the Health Services and Mental Health Administration, under contracts HSM 110-HMO-63(2) and HSM 110-72-420.

Address for reprints: Marilyn Bergner, Ph.D., Department of Health Services, SC-37, University of Washington, Seattle, WA 98195.

0025-7079/81/0800/0787/$01.45 © J. B. Lippincott Co.

change and comparative differences, attention was given to detection of low-level sickness impacts.

With the shift in emphasis in the developed countries from the curing of disease to minimizing the impact of illness on everyday activities, efficacy and efficiency of care cannot be judged by morbidity or mortality rates. Rather, estimates of the actual performance of activities are needed to provide a relevant and sensitive indicator for evaluating medical care, assessing needs and determining the allocation of resources.

We hypothesized that a broadly based assessment of performance of daily activities would provide an acceptable measure of health care outcomes that would be reliable, appropriate and sensitive to changes over time among treatment and diagnostic groups. Its basic content and form were dictated by the use to which it was directed, by a commitment to the development of a methodologically sound measure and by a concern for the practical issues of administration and feasibility.

The SIP in its final form contains 136 statements about health-related dysfunction in twelve areas of activity. It can be administered by an interviewer in 20 to 30 minutes or can be self-administered. In completing the SIP, the subject is asked to endorse or check only those statements that he is sure describe him on a given day and are related to his health. Sample statements drawn from each category of the SIP are shown in Table 1. As will be described below, the form, instructions, number of statements and areas of activity have been modified in accord with data obtained in several field trials.

Summary of Previous Work

Initial work began in 1972 with the development of procedures to collect and evaluate statements describing sickness-related behavioral dysfunction from patients, individuals caring for patients, the apparently healthy and health care professionals.

The resultant statements were subjected to standard grouping and sorting techniques yielding 312 unique items (reduced to 136 in the final form) each describing a sickness-related behavioral change. The 312 items were grouped into areas of activity or categories and then included in a prototype Sickness Impact Profile.† This questionnaire, together with its applications, reliability testing, validation and revisions was the subject of the field trials to be described.

The strategy chosen for developing, assessing and revising the SIP was based on methodological principles that emphasized the evaluation of reliability and validity in a variety of settings, the determination of the relationship of the SIP to other measures currently in use and the evaluation of its unique contribution as an outcome measure of health status.

This strategy was operationalized and implemented through a series of field trials, each designed to address specific issues in the developmental process. The sequential properties of the overall research design were particularly valuable. They provided an opportunity to answer questions that arose in earlier administrations and progressively to revise and refine the SIP.

Sampling Strategy

Field trials of the SIP were conducted in 1973 and 1974, and were designed so that the instrument would be tested on subjects that spanned a range of type and severity of illness. Since the SIP measures the behavioral impacts of sickness in terms of dysfunction and does not assess levels of positive functioning, it was assumed that the distribution of levels of sickness or dysfunction that would be obtained with sim-

† See Bergner et al.[1] for detailed information about this process.

TABLE 1. Sickness Impact Profile Categories and Selected Items

Dimension	Category	Items Describing Behavior Related to:	Selected Items
Independent Categories	SR	Sleep and rest	I sit during much of the day I sleep or nap during the day
	E	Eating	I am eating no food at all, nutrition is taken through tubes or intravenous fluids I am eating special or different food
	W	Work	I am not working at all I often act irritable toward my work associates
	HM	Home management	I am not doing any of the maintenance or repair work around the house that I usually do I am not doing heavy work around the house
	RP	Recreation and pastimes	I am going out for entertainment less I am not doing any of my usual physical recreation or activities
I. Physical	A	Ambulation	I walk shorter distances or stop to rest often I do not walk at all
	M	Mobility	I stay within one room I stay away from home only for brief periods of time
	BCM	Body care and movement	I do not bathe myself at all, but am bathed by someone else I am very clumsy in body movements
II. Psychosocial	SI	Social interaction	I am doing fewer social activities with groups of people I isolate myself as much as I can from the rest of the family
	AB	Alertness behavior	I have difficulty reasoning and solving problems, for example, making plans, making decisions, learning new things I sometimes behave as if I were confused or disoriented in place or time, for example, where I am, who is around, directions, what day it is
	EB	Emotional behavior	I laugh or cry suddenly I act irritable and impatient with myself, for example, talk badly about myself, swear at myself, blame myself for things that happen
	C	Communication	I am having trouble writing or typing I do not speak clearly when I am under stress

ple random sampling would approximate the J-curve reported for other "deviant" behaviors.[2] For early development and revision purposes, it was considered important to sample subjects who could be expected to respond to SIP items. Thus, a sampling strategy was devised that avoided simple random sampling. Rather, it required the selection of purposive samples of subjects weighted towards the sick or dysfunctional and stressed the specific and cumulative data needed to address the unique and general questions of reliability, validity, applicability, and feasibility of the SIP.

Subsequently, to assure that the final selection of SIP items, scoring methodology and format were based on data that, as far as possible, covered the range that could be expected to be encountered in the actual use of the instrument, the SIP was administered to a large random sample.

Feasibility Test of a Prototype SIP

The 1973 pilot study was aimed at providing preliminary assessments of reliability, validity and ease of administration. Two hundred and forty-six subjects (outpatients, inpatients, home care patients, walk-in clinic patients and nonpatients) completed SIPs in this field trial. In addition, a scoring method was developed and tested. This method relied on item scale values obtained from 25 judges who rated each SIP item on a 15-point scale of dysfunction that ranged from minimally dysfunctional to maximally dysfunctional. The scaling procedure and its validation is reported elsewhere.[3]

An overall SIP per cent score may be obtained by summing the scale values of all items endorsed in the entire SIP, dividing that sum by the sum of the values of all the items in the SIP and multiplying the obtained quotient by 100. Scores for each category are calculated in a like manner. That is, the scale values of all items endorsed within a category are summed, divided by the sum of the values of all items in the particular category and multiplied by 100. The scoring method was validated against ratings of dysfunction made by groups of judges (not including the 25 mentioned above) who based their ratings on the responses to unscored SIPs.[1]

The 1973 pilot study data were analyzed for purposes of item revision. A statistical analysis was used to revise and shorten the prototype instrument. Items that were relatively independent and accounted for most of the subject variance, items that were insufficiently tested and/or items that were substantively important were retained in a revised SIP for further testing. Thus, the number of items in the instrument was reduced from 312 to 189 and the wording of remaining items was refined. This revised SIP was statistically pretested by rescoring the 1973 pilot study data to assure comparability with the original measure.[‡]

Reliability and Discriminant Validity Test of the Revised SIP

The 1974 field trial was designed to provide a comprehensive test of the reliability of the SIP, a preliminary assessment of validity, a preliminary test of self-administration and a broad assessment of the revised SIP. Special attention was paid to obtaining a sample that would respond to items describing dysfunction in communication, ambulation and intellectual functioning, since the 1973 field trial had not provided sufficient responses in these categories to permit dependable item analysis.

The design utilized four subsamples of subjects covering a range of sickness or dysfunction: rehabilitation medicine outpatients and inpatients, speech pathology inpatients, outpatients with chronic health problems and a group of enrollees in a prepaid health plan who had participated in a 20-year longitudinal study and who were not ill at the time.

The test–retest reliability of the SIP was investigated using different interviewers, different forms, different administration procedures and a variety of subjects who differed in type and severity of dysfunction. Overall, the reliability of the SIP in terms of score was high (r = 0.75–0.92) and reliability in terms of items checked was moderate (r = 0.45–0.60). Reliability did not appear to be significantly affected by the variables examined, which suggests that the SIP is potentially useful for measuring dysfunction under a variety of administrative conditions and with a variety of subjects.[4,5]

[‡] In order to distinguish the three versions of the SIP that are discussed in this article, the original 312-item SIP is referred to as the prototype SIP, the 189-item revision as the revised SIP, and the 136-item SIP as the final SIP.

Validity was examined by analyzing the relationships between SIP scores and three types of measures: one based on subject self-assessment, one on clinician assessment and one on the subject's score on some other assessment instrument. SIP scores discriminated among subsamples, and the correlations between each criterion measure and SIP scores provided evidence for the validity of the SIP.[6]

The time required for completion of both interviewer-administered and self-administered SIPs, and the cost per interview, was found to be within acceptable limits for questionnaire administration.

Data from the 1974 field trial were also used in making the second revision of the SIP. The data were analyzed to determine the interrelationships among items, the relationships of items to category and overall scores and to the various criterion measures, the clarity of instructions, the reliability and clarity of items, and the discriminative capability of items. Again, caution was exercised in revising or eliminating items because the data had been obtained from purposive samples.

1976 Survey and Clinical Test of the SIP

The 1976 field trial had three basic aims: determining the final content, format and scoring of the SIP; providing a broad assessment of the discriminant, convergent and clinical validity of the SIP; and comparing the reliability and validity of alternative administrative procedures.

The Sample

To assure that the final selection of SIP items, scoring methodology and format were based on data that, as far as possible, covered the range that could be expected to be encountered in the actual use of the instrument, the administration of the SIP to a large stratified random sample of members of a prepaid group practice was a major component of the field trial. In addition, this sample was considered sufficiently diverse to include variations in illness level and sociodemographic characteristics that could affect response patterns or sickness levels.

The sampling design for the random sample was developed to provide an equal number of subjects in each of twelve sampling strata that took into consideration sex, age and type of membership in the prepaid group practice. The 696 respondents who completed the SIP represent 80 per cent of all random sample subjects contacted.

To assure an adequate frequency of response to SIP items so that final item analysis would be possible, a sample of subjects who considered themselves sick was also interviewed. This sample (known as the Quota Sample) was obtained from the patients of a family medicine clinic. All patients who had an appointment for something other than a well adult or obstetrical examination were sampled. A total of 199 subjects completed the SIP, representing 77 per cent of those sampled.

It should be noted that the sampling plan for the 1976 field trial followed the same general sampling strategy that had been employed throughout the development of the SIP. This overall strategy was aimed at testing as broad a range as possible of subjects with sickness-related dysfunction. This breadth was not obtained with one sample; instead, a series of trials that contained various samples cumulatively provided the broad range of subjects that was required for adequate testing. This succession of testing, purposely seeking increasingly less severely dysfunctional subjects can be seen in Figure 1. These graphs show that each sample on which the SIP was tested provided an increasingly higher proportion of subjects with low SIP scores. Thus, a typical J-curve distribution of sickness-related dysfunction was obtained by the time the 1976 field trial was completed, assuring adequate testing of the in-

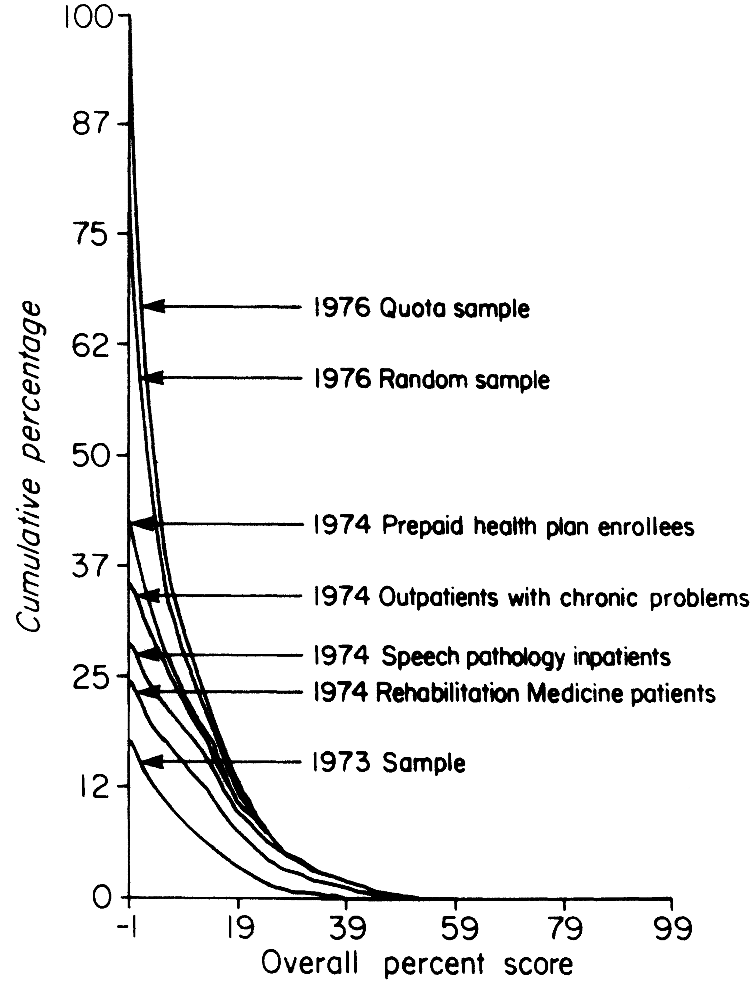

FIG. 1. Cumulative percentage contribution of each SIP sample to the SIP score distribution of the cumulative sample (N = 1,108).

strument in the important low severity range.

Reliability

Because the 1973 and 1974 field trials included extensive tests of general reliability in terms of reproducibility and internal consistency, the 1976 field trial contained only sufficient reliability testing to assure that previous levels were maintained. (Fifty three subjects were administered two SIPs within a 24-hour period.) Cronbach's alpha was calculated to assess internal consistency. As can be seen in Table 2, reliability of the SIP is comparable across all field trials. As was expected, reliability in terms of score is high; in terms of item agreement only moderate. This suggests that though subjects change the specific items they respond to within a 24-hour period, the combination of items checked on the two occasions are sufficiently similar in scale value to provide similar overall and category scores. It should be noted that the test used to calculate item agreement was very conservative. It counted as agreements only those items that had a *positive* response in both SIPs and as disagreements those items that had a positive response in only one of the SIPs. It disregarded agreement in the form of negative responses in both SIPs:

$$\text{agreement \%} = \frac{\text{no. of positive agreements}}{\text{no. of positive agreements plus number of disagreements}}$$

The 1976 field trial provided an opportunity to compare the reliability of three types of administration of the SIP: an interviewer administration (I), an interviewer-delivered self-administration (ID), and a mail-delivered self-administration (MD).

As in the 1974 field trial, high levels of test–retest reliability, evaluated in terms of score correlations, were demonstrated for Is and IDs (no retests on MDs could be obtained) and analyses of variance showed no difference in overall mean scores between administrative types. Internal consistency (Cronbach's alpha, Table 3) was high for both Is and IDs but substantially lower for MDs.

To further assess the comparability of administration types, the relationship of SIP score to self-assessments of sickness and dysfunction, clinician assessments of dysfunction and an index of disability derived from the National Health Interview Survey restricted activity days questions (NHIS) was determined. Although some differences were noted, no single administration type consistently displayed stronger relationships to criterion variables than did any other. Lower correlations were noted for MD overall SIP score and the NHIS index (Table 3). (Separate correlations between category scores and self-assessments of dysfunction for MDs were markedly lower than those of Is and IDs for categories SI [Social Interaction], E [Eating], HM [Home Management], M [Mobility], and BCM [Body Care and Movement.].)

In summary, it appears that mail-delivered SIPs may not provide data comparable to that obtained by the other two types of administration. Both types of self-administered SIP, however, provided somewhat higher mean scores and the interviewer-delivered self-administered form of the SIP showed consistently higher correlations with the other measures of

TABLE 2. Reliability Summary of the SIP Across All Field Trials

	1973 Field Trial	1974 Field Trial	1976 Field Trial
Reproducibility			
Overall score	0.88	0.88	0.92
Category items	0.56	0.50	0.50
Internal consistency			
Cronbach's alpha	NA	0.97	0.94

NA: not applicable.

TABLE 3. Reliability Summary for Interviewer Administered, Interviewer-delivered Self-Administered and Mail-delivered Self-Administered SIPs

	I	ID	MD
Test–retest reliability*	0.97	0.87	NA
Internal consistency	0.94	0.94	0.81
Mean and standard deviation of SIP score	2.6 (4.5)	3.6 (5.3)	3.0 (4.2)
Correlation of SIP score with other measures			
Self-assessment of dysfunction	0.64	0.74	0.48
Self-assessment of sickness	0.55	0.67	0.38
NHIS†	0.57	0.60	0.05

I: Interviewer administration.
ID: Interviewer-delivered self-administration.
MD: Mail-delivered self-administration.
NA: Not applicable.

* The difference in test–retest reliability between the Is and IDs is statistically significant ($p < 0.01$).

† National Health Interview Survey Index of Activity Limitation, Work Loss and Bed Days.

dysfunction and sickness than interviewer-administered SIPs. These data suggest that self-administered forms may be more valid than interviewer-administered forms when accompanied by a method of administration that assures comprehension of and adherence to SIP instructions, and conveys a sense of importance of the task. A trained interviewer who reads instructions and answers questions before the SIP is completed by the subject may be the best assurance of reliable and valid SIP data. If mail-delivered SIPs must be used, careful follow-up and monitoring is necessary to assure and assess reliability and validity.

Validity

The validity of the SIP depends on demonstrating the relationship between sickness impacts and behavioral dysfunction. In each field trial an attempt was made to determine the relationship between independent measures of sickness and of dysfunction.[1,6] Subjects were asked to rate their overall level of dysfunction and overall level of sickness. High correlations between these two measures were obtained from subjects in all SIP field trials, providing evidence for the validity of the sickness impact–dysfunction relationship.

During 1973, the concept of dysfunction was employed in scaling the individual SIP items and in rating SIP protocols obtained from field trial subjects. Since the SIP items were derived from empirically obtained statements describing sickness-related behaviors, the strong relationship between judgments and SIP scores based on item scale values provided further evidence of construct validity.[1]

In the 1974 field trial, preliminary estimates of the validity of the SIP were obtained by examining the relationship between the SIP and self-assessments of dysfunction, between the SIP and other measures of dysfunction, including the Activities of Daily Living Index,[7] and between the SIP and selected questions from the National Health Interview Survey.[8] In general, the relationships between overall SIP scores and the criterion measures were high. Although these data provided preliminary evidence of SIP validity, additional and more refined criterion measures were clearly needed.

In the 1976 field trial, subject and clinician assessments of health status were obtained and the relationships between SIP scores and these measures were examined. In addition, the relationship between SIP scores and clinical measures of patient progress was determined.

On the basis of previous work,[6] we assumed that the strength of the relationship between the SIP and other measures of health status was a function of the similarity of the construct being measured and the similarity of the method of measurement. Therefore, a series of hypotheses concerning these relationships was generated. We hypothesized that SIP score would be more related to those criterion measures reflecting subject perceptions than to other criteria. Specifically, we hypothesized that SIP scores would be most related to subject self-assessments of dysfunction. This hypothesis was based on similarity of the construct measured and the method used in obtaining the measurements. Subjects were administered a given category of the SIP and instructed to respond to those statements that described them and were related to their health. Then, they were asked to rate their relative level of dysfunction in that area of activity on a seven-point scale. Finally, subjects were asked to rate their overall level of dysfunction. Thus, the subject made both responses within a common area of activity and both measures were designed to tap the *same* construct of dysfunction.

We hypothesized further that the next highest relationship would be between SIP score and self-assessment of sickness. Self-assessments of sickness, like self-assessments of dysfunction, are subject-reported perceptions, but the construct of sickness and the method used to measure it differed somewhat from those employed in the SIP. Although sickness is an integral part of the conceptualization underlying the SIP, it does not directly tap dysfunction and, therefore, was not expected to be as strongly related to the assessment of dysfunction.

SIP scores were hypothesized to be less related to the NHIS index than to the self-assessments of dysfunction or sickness, because this index differs from the SIP in two aspects. It refers to a 14-day period rather than the 1-day period of the SIP, and it measures restricted activity days in a general fashion. Nonetheless, it reflects the same underlying construct of dysfunction and is a self-perceived report of limitation of activity.

We assumed that the SIP would be less related to measures of sickness obtained from sources other than the subject. Thus, we hypothesized that the SIP would be least related to clinician ratings. Clinicians rated both dysfunction and sickness levels of their patients. They were asked to rate their patients' level of dysfunction in each of the SIP categories, keeping in mind the scope covered by the SIP items in that category. Then, the clinician was asked to make an overall rating of dysfunction and an overall rating of sickness. Further, we hypothesized that because of the common construct of dysfunction, SIP score would be more related to clinician ratings of dysfunction than to clinician ratings of sickness.

Analysis of the 1976 field trial data confirms these hypotheses. The correlation between SIP score and self-assessment of dysfunction is 0.69; between SIP score and self-assessment of sickness is 0.63; between SIP score and the NHIS index is 0.55; between SIP score and the clinician assessment of dysfunction is 0.50; and between SIP score and the clinician assessment of sickness is 0.40. These data across all field trials are summarized in Table 4.

The relationships between the SIP and each of the criterion variables were further analyzed by the multitrait–multimethod methodology developed by Campbell and Fiske[9] and by multiple regression techniques. The multitrait–multimethod technique assesses convergent and discriminant validity by examining the relative effect of the method of measurement and the construct or trait being measured on the correlations among measures.

A summary of the multitrait–multimethod matrix is presented in Table 5. Examination of the first column of the table shows that the reproducibility of category scores and overall scores is markedly higher than any of the correlations among different category scores. The relatively low correlations among category scores (see rows 2 and 3) assures minimal redundancy; the higher correlation of category scores (see rows 4 and 5) to overall scores assures the importance of each category to the total instrument. Examination of the first and last rows shows that the reproducibility of SIP scores is higher than the reproducibility of other measures of sickness or dysfunction, and that SIP scores are more highly related to those criterion

TABLE 4. Validity Summary of the SIP Across All Field Trials

Criterion	1973 Field Trial	1974 Field Trial	1976 Field Trial
Protocol judgments	0.85	NA	NA
Self-assessment			
sickness	NA	0.54	0.63
dysfunction	NA	0.52	0.69
Clinician assessments			
sickness	NA	0.30	0.40
dysfunction	NA	0.49	0.50
Other instruments			
NHIS*	NA	0.61	0.55
ADL†	NA	0.46	NA

NA: Not applicable.

* National Health Interview Survey Index of Activity Limitation, Work Loss and Bed Days.

† Activity of Daily Living Index

TABLE 5. Summary of a Multitrait–Multimethod Matrix for the Sickness Impact Profile*

	SIP	SAD	SAS	NHIS	SAD and SIP	SAS and SIP	NHIS and SIP	CROS and SIP	CROD and SIP
Mean correlation of each category with itself, Time 1 and Time 2	0.82 ± 0.08	0.76 ± 0.12	NA	NA	0.66 ± 0.06	NA	NA	NA	0.41 ± 0.11
Mean correlation of each category with every other category, Time 1	0.32 ± 0.19	0.41 ± 0.20	NA	NA	0.38 ± 0.09	NA	NA	NA	0.27 ± 0.14
Mean correlation of each category with every other category, Time 2	0.40 ± 0.21	0.63 ± 0.13	NA	NA	0.35 ± 0.11	NA	NA	NA	0.27 ± 0.14
Mean correlation of each category woth overall score, Time 1	0.60 ± 0.16	0.56 ± 0.13	NA	NA	0.44 ± 0.09	0.40 ± 0.11	0.35 ± 0.11	0.26 ± 0.08	0.32 ± 0.09
Mean correlation of each category with overall score, Time 2	0.66 ± 0.17	0.67 ± 0.10	NA	NA	0.59 ± 0.07	NA	NA	NA	0.42 ± 0.12
Correlation of overall score, Time 1 and Time 2	0.92	0.82	0.86	0.85	0.69	0.63	0.55	0.40	0.50

SAD: self-assessment of dysfunction.
SAS: self-assessment of sickness.
NHIS: National Health Interview Survey Index of Activity Limitation, Work Loss and Bed Days.
CROS: clinician rating of sickness.
CROD: clinician rating of dysfunction.
NA: not applicable.

* This summarizes a complete multimethod-multitrait matrix in which each of the measures is correlated with every other measure for each category of the SIP, following the methodology described by Campbell and Fiske.[9] The data in the above table presents the mean and standard deviation of all obtained correlations. The original matrices may be obtained from the authors.

measures that were, a priori, considered to be most reflective of the construct of sickness and the methodology employed in the SIP.

To further test the convergent and discriminant validity of the SIP as hypothesized above, a multiple regression analysis was undertaken. This analysis was aimed at determining the amount of variance explained by SIP category scores in each of the criterion measures used across all field trials. Results of these analyses are shown in Table 6. The SIP explains less of the variance in measures of sickness (SAS and Speech Pathology ratings) than in measures of dysfunction (SAD, CROD, ADL). These data provide confirmation of the multimethod–multitrait analysis.

TABLE 6. Per Cent of Variance in Criterion Measures Explained by SIP Category Scores in Stepwise Multiple Regression

	1973	1974	1976 Random	1976 Quota
Protocol judgments	0.79			
SAD		0.41	0.51	0.56
SAS		0.37	0.45	0.48
NHIS		0.45	0.39	0.52
CROD		0.59*		
ADL		0.60		
Speech pathology ratings		0.30		

SAD: self-assessment of dysfunction.
SAS: self-assessment of sickness.
NHIS: National Health Interview Survey Index of Activity Limitation, Work Loss and Bed Days.
CROD: clinician rating of dysfunction.
ADL: Activity of Daily Living Index.
* Obtained only for the outpatients with chronic problems.

Clinical Validity

Another group of criteria against which a health status measure should be validated consists of objective clinical data that are characteristically used to follow the progress of patients with specific diagnostic conditions.

The SIP has been designed to be applicable to and to provide information about the sickness-related dysfunctions of individuals as well as groups. This faculty, if demonstrated, should be of particular importance to clinicians in evaluating alternative modes of treatment, assessing progress of a particular patient and providing information about diagnosis and patient management. The test of this faculty of the SIP involves 1) assessment of the relationship between the SIP and existing clinical measures; and 2) determination of whether the SIP provides additional information not provided by the existing clinical measures. The former is important if clinicians are to be assured that the information obtained from the SIP is consonant with more traditional data obtained on patients. The latter is important in order to assess and specify the types and range of supplemental information that can be provided by the SIP. The relevance of the clinical validation of the SIP to clinical medicine is discussed in a forthcoming article.[10]

Demonstration of clinical validity requires the selection of clinical measures that are generally believed to be related to patient function. Three disease categories were chosen for which clinicians concurred that there are reliable clinical measures that parallel the patient's functional health status. The disease categories were total hip replacement, hyperthyroidism, and rheumatoid arthritis.

Although specific tests, time intervals and conditions differed for each group, the general format for study of each of these diagnostic groups was as follows: 1) each diagnostic group contained fifteen patients; 2) patients were measured at least three times during the study period; 3) follow-up times were specified in advance and procedures developed to assure the timely administration of SIPs and collection of clinical data; and 4) clinical measures and SIPs were obtained within a 24-hour period. Thus, an estimate of variability or response error could be obtained.

The data obtained for each of the diagnostic groups are shown in Table 7.

To assess the relationship between the SIP and the clinical measures, correlations between them were obtained. As can be seen in Table 8, these correlations are moderate (r = 0.41) to high (r = −0.84). On the basis of discussions with clinicians, several specific hypotheses were tested concerning the relationship between groups of SIP categories and the clinical measures.

Since the Harris Analysis of Hip Function measures only physical dysfunction and pain, it was hypothesized that scores based on a combination of SIP categories that describe physical dysfunction (Dimension I) would be more highly correlated with Harris Analysis of Hip Function than would overall SIP scores or scores based on a combination of categories describing phychosocial dysfunction (Dimension II). In general, the data support this hypothesis. Though there is little difference between the SIP overall score correlations with the Harris Analysis of Hip function (r = 0.81) and the Dimension I Score correlation with the Harris Analysis of Hip Function (−0.84), the Dimension II Score correlation with the Harris Analysis of Hip Function is considerably lower (r = 0.61) than either of the former.

The clinical picture of hyperthyroidism suggests substantial impact on the psychosocial areas. Therefore, we hypothesized that scores on Dimension II (Psychosocial Dimension) would be more highly correlated with adjusted T4 (thyroid hormone) than would overall SIP scores and Dimension I (Physical Dimension) scores. This hypothesis is only partly supported by the data. Dimension II scores are more highly correlated with adjusted T4 (r = 0.35) than are Dimension I scores (r = 0.21), but not more correlated than are overall SIP scores (r = 0.41). Since the overall SIP contains categories not included in either the Physical or Psychosocial Dimension, we assumed that it was these independent categories that accounted for this higher correlation with overall SIP score. Indeed, the category Sleep and Rest showed the best relationship to adjusted T4 level.

In view of the nature of rheumatoid arthritis and the clinical criteria that were examined, several hypotheses were generated to guide the analysis of the data. First, it was hypothesized that Dimension I scores would be more highly correlated with grip strength, walking time, number of painful joints and number of swollen joints than Dimension II scores. This hypothesis was based on the notion that these clinical measures would be more accurately reflected in physical function than in psychosocial function, even though the

TABLE 7. Data Obtained on Patients Included in the Clinical Validation of the SIP at Initial and Follow-Up Visits

Diagnostic Group	Data Obtained
Hip replacement patients (N = 15)	Harris Analysis of Hip Function* Self-assessment of dysfunction Self-assessment of sickness Clinician assessment of dysfunction SIP
Hyperthyroid patients (N = 14)	Adjusted T_4† Pulse Self-assessment of dysfunction Self-assessment of sickness SIP
Arthritic patients (N = 15)	Activity Index‡ Self-assessment of dysfunction Self-assessment of sickness SIP

* An assessment of patients who have undergone hip replacement. A high score on this test indicates better hip function than does a low score.[11]

† A hormonal measure of thyroid function.

‡ An index developed by Haastaja[12] that combines weighted values for duration of morning stiffness, grip strength, sedimentation rate and joint involvement.

clinical literature concerning arthritics indicates that the psychosocial areas may be seriously affected. Second, it was hypothesized that SIP scores would be minimally correlated with erythrocyte sedimentation rate (ESR) and hematocrit (HCT). This hypothesis was based on evidence in the medical literature that ESR and HCT do not accurately reflect the functional impact of the disease on the patient. Third, it was hypothesized that Dimension II scores would be more highly correlated with patient's assessment of pain, ease of movement and "how they feel" than Dimension I scores. These criteria seemed more likely to be reflected accurately in psychosocial functioning, since they involved patient assessment of impact.

In general, the hypotheses are supported. Dimension I score correlations with grip strength, walking time, number of painful joints and number of swollen joints are higher than Dimension II correlations with these criteria. In addition, the correlations of the criteria are slightly higher with Dimension I scores than they are with overall SIP score. This suggests that the categories not included in either dimension score (E, HM, W, SR, RP) and the Physical Dimension categories are more sensitive than the categories in the Psychosocial Dimension to these criterion measures. SIP overall score and erythrocyte sedimentation rate is uncorrelated; and SIP overall score and hematocrit has a low correlation (r = −0.25).

Descriptive Validity

A measuring instrument may evidence construct, convergent and discriminant validity and yet have little capacity to describe the qualitative differences and similarities in particular samples of subjects, or in the same subjects studied longitudinally. It is important, therefore, to assess the instrument's capacity to describe and delineate samples of subjects that differ in mean score and samples that are similar in mean score. With respect to the SIP as an instrument for measuring health status, it seems crucial to know the extent to which dimension and category scores and item-checking patterns provide a useful and meaningful qualitative description of different samples and types of subjects. Pattern and profile analyses of SIP sensitivity have been performed on the individual and the diagnostic group data as further tests of the validity of the SIP.

A detailed description of the application of pattern and profile analysis[13,14] to the SIP will be discussed in a subsequent article.[15] The approach that is most appropriate to the small diagnostic samples employs a modification of the methodology suggested by Cronbach and Gleser.[13] The profiles of SIP category scores obtained for each diagnostic group at each point in time were assessed in terms of mean differences (elevation), variability differences (scatter) and pattern differences (shape).

A graph of SIP category scores for specific patients provides a profile of the

TABLE 8. Correlations of SIP Scores and Clinical Measures

Clinical Measures	Correlation with Overall SIP Score	Correlation with Categories that Measure Physical Dysfunction	Correlation with Categories that Measure Psychosocial Dysfunction
Harris Analysis of Hip Function*	−0.81	0.84	0.61
Adjusted T_4†	0.41	0.21	0.35
Activity Index‡	0.66	0.66	0.56

* An assessment of patients who have undergone hip replacement. A high score on this test indicates better hip function than does a low score.[11]

† A hormonal measure of thyroid function.

‡ An index developed by Haastaja[12] that combines weighted values for duration of morning stiffness, grip strength, sedimentation rate and joint involvement.

dysfunctions experienced by these patients. If patients with a particular diagnosis exhibit similar SIP scores, a consistent profile of dysfunction for that diagnosis will emerge.

The profiles of hip replacement patients (Fig. 2, 3 and 4) are provided as illustration. Hip replacement patients show a consistent pattern of dysfunction across all patients and all administrations of the SIP. This pattern is characterized by minimal dysfunction in the pychosocial areas and substantial dysfunction in the physical areas. Though the amount of dysfunction differs over time, the pattern of dysfunction appears to persist.

With respect to elevation, or differences in mean score across categories (Fig. 2), t-tests of pairs of profile means confirmed that the severity of dysfunction differed significantly between the four points in time.

With respect to scatter, or the variability among mean category scores (Fig. 3), the profiles show a significant difference between times ($p \leq 0.05$). With respect to shape, or pattern across time controlling for elevation and scatter (Fig. 4), the shape at Times 1 and 3, Times 1 and 4, and Times 2 and 3 are more highly correlated than the other comparisons. These data point to a difference in the shape of the Time 2 profile from the others. Time 2 SIPs were completed while the patient was hospitalized and most dysfunctional. The score difference (elevation) signals the greater dysfunction; the shape difference signals a pattern of dysfunction that may be characteristic of the hospitalized patient.

Hyperthyroid patients, like hip replacement patients, exhibit a characteristic profile of sickness impacts that can be readily discerned. The group profile shows moderate dysfunction in the psychosocial categories of the SIP and substantial dysfunction in the independent categories, notably SR (Sleep and Rest), HM (Household Management), and RP (Recreation

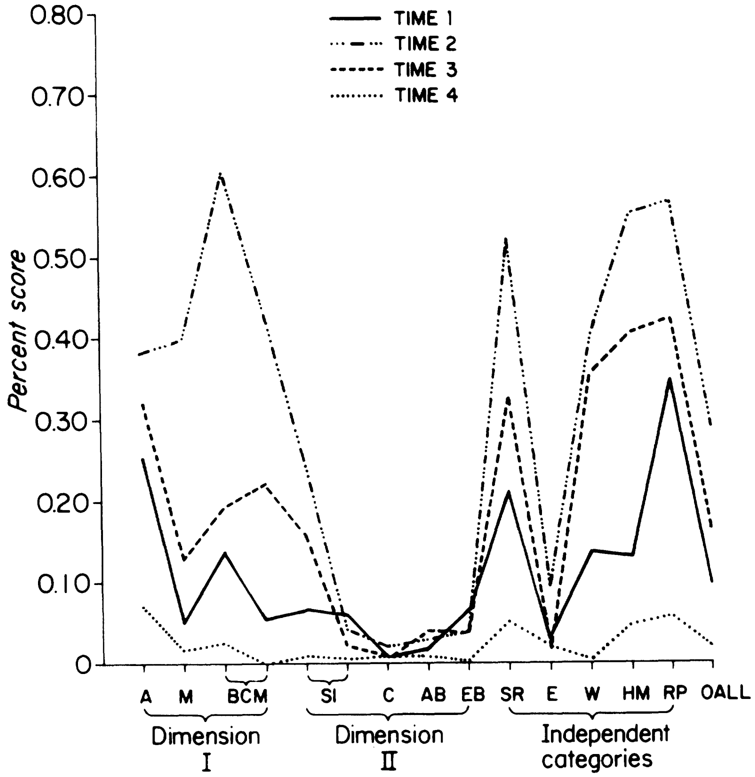

FIG. 2. Elevation of SIP category scores for total hip replacement patients at each follow-up visit.

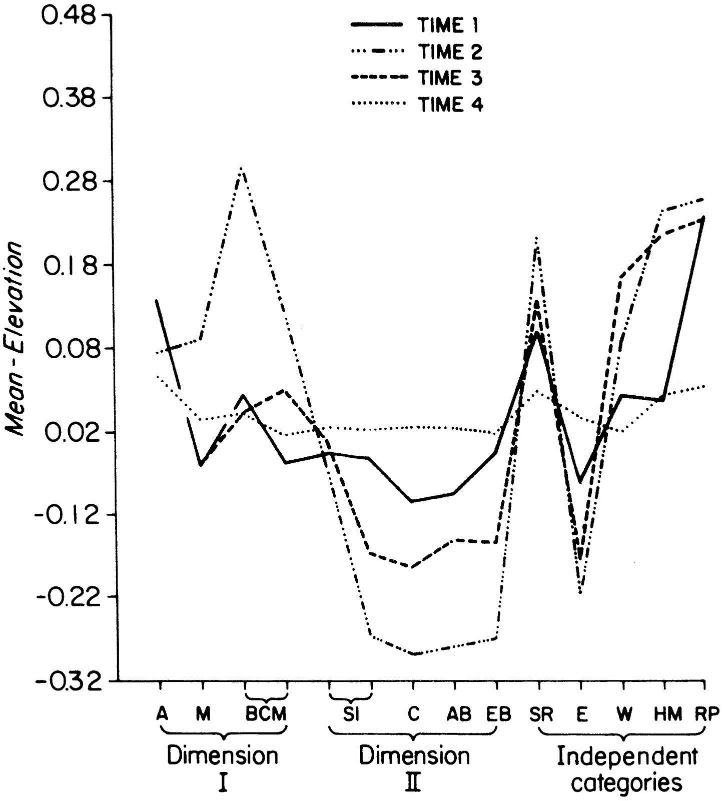

FIG. 3. Scatter of SIP category scores for total hip replacement patients at each follow-up visit.

and Pastimes). This is particularly apparent at the initial SIP administration. In addition, significant differences between mean SIP score at Time 1 and Time 4 were found, as was found for adjusted T4 levels.

Examination of the profiles of SIP category scores for rheumatoid arthritic patients provides a picture of a disease with impacts that are idiosyncratic to each patient. In contrast to the hip replacement and hyperthyroid patients, each arthritic patient has a distinct SIP profile that looks like test–retest reliability profiles in that it does not change over time and seems unaffected by changes in treatment.

The cluster analysis approach[14] to pattern analysis was also applied to each diagnostic group across times, allowing for the definition of a cluster of categories that consistently differentiated among groups of patients and for each group of patients among the different points in time.

The study of the three diagnostic groups with regard to validity and sensitivity supports the value of the SIP as a measure of health status. The findings are consistent with clinical observations while providing information that in some cases is new, and in others is a complement that highlights clinical observations that may have been ignored or deemphasized.

Revision and Statistical Pretest of the Final SIP

The final revision of the SIP was based on data from the 1976 field trial and a Cumulative Sample of subjects' responses in the 1973, 1974 and 1976 field trials. SIP data on some 2,000 subjects from an Alabama study[16] were also examined. The following was also taken into consideration: 1) a consumer validation of the original severity of dysfunction scaling of

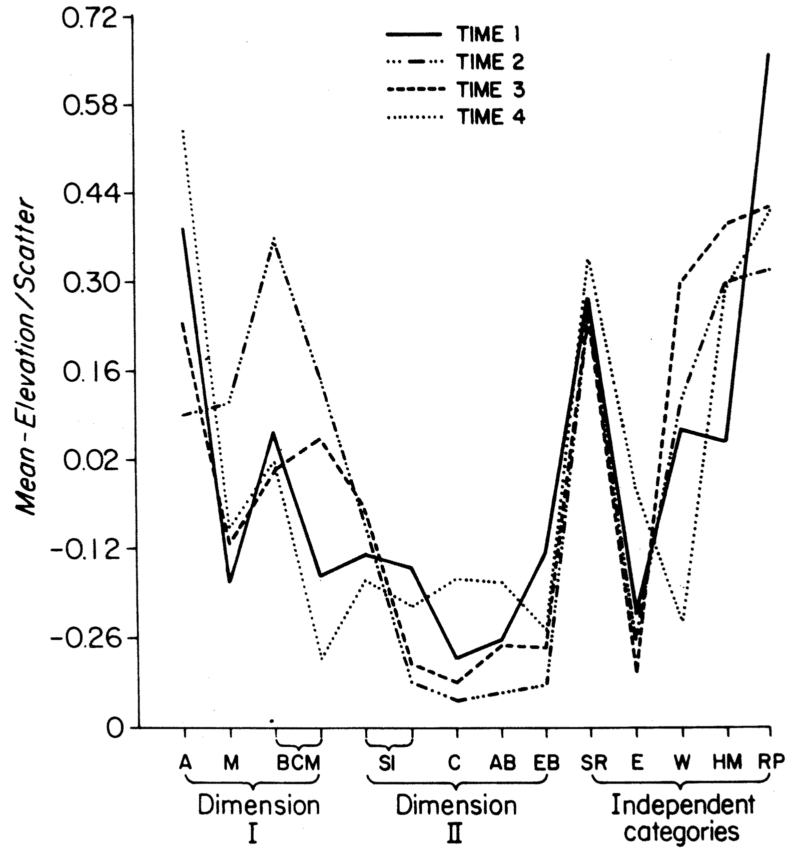

FIG. 4. Shape of SIP category scores for total hip replacement patients at each follow-up visit.

items[3]; 2) an analysis of the relationships among the SIP categories; 3) the discriminative capacity of the final SIP items; and 4) an examination of methods of scoring and degrees of scoring aggregation.§

Category Analysis

Since the SIP is designed to describe various kinds of dysfunctions in different areas of activity, it contains groups or categories of items that are interrelated. It is necessary to study the magnitude of these interrelationships and to determine where there is such complete overlap as to approach redundancy, and where there is important reflection of a basic dysfunction across several areas of activity. In the first instance, categories should be combined or eliminated; in the second, the dysfunction score should reflect this generalization of impact.

Validity analyses have shown that while all SIP categories are not required to account for variance among subjects in each subsample and on each criterion measure, each category is important in one or more instances. Also, the more dysfunctional the sample, the more categories are responded to, and the higher the intercorrelation of category scores and the correlation of each category with overall SIP score. At the present time, it is difficult, if not impossible to predict, a priori, which categories will be most important for a particular sample.

Assuming that some categories should be eliminated or combined, specific hypotheses were evaluated statistically and conceptually. We hypothesized that the category of items concerned with eat-

§ During each of the field trials, two lengths of the SIP were tested. One contained only those items that were statistically discriminative, the other contained those items plus items that were thought to have been insufficiently tested or that were of descriptive importance.

ing and taking nutrition did not make a statistically significant contribution to the SIP. The statistical analyses showed that this category did not account for a significant amount of variance among subjects. However, consideration of the application of the SIP in clinical and program evaluation settings indicated that this category made a substantive contribution to the descriptive capacity of the instrument. On this basis, this category of items was retained.

In the original development of the SIP, items dealing with work inside and outside the home were judged to represent separate categories of behavior. On the basis of subsequent responses to the two categories, we hypothesized that work both inside and outside the home could be measured on a single continuum with no loss in statistical sensitivity. Various statistical analyses did not support this hypothesis and the combination appeared, in fact, to distort the obtained results. Retention of the two as separate categories in the SIP provides a more integrated and sensitive instrument for use with all types of samples.

High intercategory correlations were obtained between several pairs of categories. These intercorrelations suggested that several possible combinations of categories could be made to reduce redundancy. Further statistical analyses of these combinations consistently supported the combination of Movement of the Body (M) and Personal Hygiene (BC). Thus, category BCM (Body Care and Movement) was adopted. Though the combination of Social Interaction (SI) and Family Interaction (FI) was not as consistently supported by the category score analyses, the subsequent item analysis provided conclusive support for the eventual combination of the two categories into a single category, Social Interaction (SI), that contained items describing dysfunctional behavior in family interaction as well as in more generalized social interactions.

Item Analysis

Item analyses were conducted to assess 1) the relationships among all items; 2) the relationships between SIP items and SIP category scores; 3) the relationships between SIP items and a number of criterion variables; 4) the differences among sampling strata in terms of the number of times each item was checked; and 5) the reliability of each item.

The relationships between items and category scores and between items and criterion variables were examined in correlational and item-checking pattern analyses.[||] Also examined were results of stepwise multiple regression, interaction detection (MAID) and item cluster analyses.

Coefficients of association and MAID permitted an assessment of redundancy. For example, an item might have a high coefficient of association with another or a combination of other items, but be differentially predictive depending on whether the other item or combination of items was checked.

Tentative conclusions regarding item disposition were drawn from the above data. These conclusions were validated or modified by review of the following: 1) cluster analysis which identified those items most highly interrelated within clusters and at the same time independent of items included in other clusters; 2) the frequency with which an item was checked across the various demographic strata and subsamples, which indicated whether it had unique descriptive value; 3) the correlation of an item with various criterion variables, which provided a means of deter-

[||] The inter-item correlation matrices were used in conjunction with a modified Kulzinski coefficient of association between each pair of items. The coefficient takes account only of pairs of items that contain a positive response, and omits consideration of occasions when neither item is responded to positively. This provides a better measure of covariation or overlap than the Phi correlation that was used in multiple regression.[17]

mining the item's validity in terms of other estimates of health status; 4) the Agreement Per Cent, which provided test–retest reliability estimates of each item; 5) comments collected during field trials about wording and administrative difficulties relative to specific items, which suggested appropriate item revisions; 6) items for which consensus in scaling did not meet the criterion, which suggested that they should be revised and rescaled.

During the review process, 53 items that were in the revised SIP were dropped or combined with other items. The final SIP contains 136 items in 12 categories. Three of these categories may be combined into a Physical Dimension; four others into a Psychosocial Dimension. The remaining four categories are independent and each may be scored separately. All items in all categories are included in the overall SIP score.

Items were retained in the SIP on the basis of their discriminative capacity within their particular category. Therefore, the reliability and validity of individual category scores are maintained.#

Statistical Pretest

The value of the final revision of the SIP will derive from tests of it in the field. It has, however, been possible to assess to some extent how well this shortened and modified instrument would have accounted for variance among subjects to whom former, more extensive SIPs were administered. In a statistical pretest, revised SIP scores were derived for all previous field trial subjects and a set of analyses performed:

1. Reliability in terms of internal consistency.[18]
2. Validity in terms of a multitrait–multimethod matrix using the derived SIP category and overall scores in correlations with criterion measures.
3. Validity in terms of stepwise multiple regression using derived scores to account for variance among subjects on criterion measures.
4. Comparisons of the correlations of original and derived SIP scores with demographic variables in the 1976 random sample; derivation of mean score estimates based on a reconstituted sample that weights the estimate for each stratum according to the proportions found in the prepaid health care facility.

In all these analyses the 136-item SIP did as well or better than the earlier, longer versions of the SIP. Alpha coefficients for the earlier version of the SIP and the 136-item SIP are comparable throughout all categories and overall. Correlation by category of the revised SIP and the final SIP scores with criterion measures shows that no category lost a significant amount of discriminative capacity and that the final SIP accounts for approximately the same amount of variance among subjects on criterion measures as the longer version. Means and standard deviations of derived and administered SIP scores by demographic stratum for the 1976 Random Sample showed that derived scores are consistently, though very slightly, higher and that standard deviations are consistently, but slightly, lower. The same relationships of scores across demographic strata are maintained. This suggests that those items not included in the final instrument were either not checked or were checked across all strata of the sample.

In summary, the same results can be expected for the final version of the SIP as have been demonstrated throughout the various field trials. Reduction of instrument length from 312 to 136 items and from 14 to 12 categories appears to maintain a breadth of assessment and discriminative power that is comparable to the original instrument.

Conclusion

The extensive testing and revision done during the development of the Sickness

# Detailed information about the revision process may be found in Gilson, et al.[18]

Impact Profile has been only briefly described above. These main findings should permit other investigators and clinicians to assess its value and relevance to their needs. Reliability has been clearly demonstrated. Construct, convergent and discriminant validity has been assessed and deemed appropriate for an instrument that seeks to measure a characteristic for which there is no criterion. Sensitivity of the instrument to different conditions or diagnoses has been tested and results obtained indicate the value of the SIP in describing similarities of groups of patients and differentiating among these groups. To our knowledge such systematic description and differentiation has not been possible heretofore.

Demonstration of the value of the SIP and further development depends now on those who choose to use it. Several large-scale studies, both in the United States and abroad, are now in progress. Clinical trials of therapy for patients with chronic lung diseases, of emergency service for cardiac arrest victims, of early exercise therapy for patients who have had myocardial infarcts and of home care for patients who are chronically ill are using the SIP as an outcome measure. The SIP is also being used to help plan services for the handicapped by administering it as part of a general survey instrument to a sample of handicapped and at-risk for handicap in a geographically defined area of London. Results of most of these studies are not yet available, but informal communication from investigators indicate that they find the SIP is feasible to administer even to the very sick, is relevant to their needs, and adds information beyond that provided by other data.

Acknowledgment

The authors wish to acknowledge the contribution of two anonymous reviewers to the completeness and clarity we may have achieved in this article.

References

1. Bergner M, Bobbitt RA, with Kressel S, Pollard WE, Gilson BS, Morris JR. The Sickness Impact Profile: conceptual formulation and methodology for the development of a health status measure. Int J Health Serv 1976;6:393.

2. Allport FH. The J-curve hypothesis of conforming behavior. J Soc Behav 1934;141:183.

3. Carter WB, Bobbitt RA, Bergner M, Gilson BS. The validation of an interval scaling: the Sickness Impact Profile. Health Serv Res 1976;(Winter):516.

4. Pollard WE, Bobbitt RA, Bergner M, Gilson BS. The Sickness Impact Profile: reliability of a health status measure. Med Care 1976;14:146.

5. Pollard WE, Bobbitt RA, Bergner M. Examination of variable errors of measurement in a survey-based social indicator. Social Indicators Research 1978;5:279.

6. Bergner M, Bobbitt RA, Pollard WE, Martin D, Gilson BS. The Sickness Impact Profile: Validation of a health status measure. Med Care 1976;14:57.

7. Katz S, Ford A, Moskovitz RW, et al. Studies of illness in the aged: the index of ADL. JAMA 1963;185:914.

8. U.S. Department of Health, Education, and Welfare. Interviewing methods in the health interview survey. (Vital and health statistics. Series 2, no. 48, 1972.)

9. Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait–multimethod matrix. Psychol Bull 1959;56:81.

10. Gilson BS, Bergner MB, Bobbitt RA, Carter WB. Clinical application of a measure of functional health status. Forthcoming.

11. Harris WH. Preliminary report of results of Harris total hip replacement. Clin Ortho 1973;95:168.

12. Haataja M. Evaluation of the activity of rheumatoid arthritis. Scand J Rheumatol Suppl 1975;4:7.

13. Cronbach LJ, Gleser GC. Assessing similarity between profiles. Psychol Bull 1953;50:456.

14. Tryon RC, Bailey DE. Cluster analysis. New York: McGraw-Hill, 1970.

15. Carter WB, Bobbitt RA, Bergner M, Gilson BS. Pattern and profile analysis of the Sickness Impact Profile. Forthcoming.

16. Miles DL. Health Care Evaluation Project: terminal project report. National Center for Health Services Research, USDHEW, 1977, photocopy.

17. Sokal RR, Sneath PHA. Principles of Numerical Taxonomy. San Francisco: WH Freeman & Co, 1963.

18. Gilson BS, Bergner M, Bobbitt RA, Carter WB. The Sickness Impact Profile: final development and testing, 1975–1978. Seattle: Department of Health Services, School of Public Health and Community Medicine, 1979.

19. Cronbach LJ: Coefficient alpha and the internal structure of tests. Psychometrika 1951;16:297.