CONCEPTUALIZATION AND MEASUREMENT OF HEALTH FOR ADULTS IN THE HEALTH INSURANCE STUDY: VOL. I, MODEL OF HEALTH AND METHODOLOGY

PREPARED UNDER A GRANT FROM THE U.S. DEPARTMENT
OF HEALTH, EDUCATION, AND WELFARE

JOHN E. WARE, JR., ROBERT H. BROOK,
ALLYSON DAVIES-AVERY, KATHLEEN N. WILLIAMS,
ANITA L. STEWART, WILLIAM H. ROGERS,
CATHY A. DONALD, SHAWN A. JOHNSTON

R-1987/1-HEW
MAY 1980

Rand
SANTA MONICA, CA. 90406

The research reported herein was performed pursuant to Grant Nos. 016B-7501-P2021, 016B-7601-P2021, and 016B-7901-P2021 from the U.S. Department of Health, Education, and Welfare, Washington, D.C.

Library of Congress Cataloging in Publication Data

Rand Corporation.
  Model for health and methodology.

  (Conceptualization and measurement of health status for adults in the health insurance study ; v. 1) ([Report] - Rand Corporation ; R-1987/1-HEW)
  Bibliography: p.
  1. Health surveys--Statistical methods. 2. Health status indicators. 3. Health surveys--United States. 4. Insurance, Health--United States. I. Ware, John E., 1942- II. United States. Dept. of Health, Education, and Welfare. III. Title. IV. Series: Conceptualization and measurement of health for adults in the health insurance study ; v. 1. V. Series: Rand Corporation. Rand report ; R-1987/1-HEW.
AS36.R3 R-1987/1 [RA409] 081s [614.4'2'028]
ISBN 0-8330-0210-4 80-7810

The Rand Publications Series: The Report is the principal publication documenting and transmitting Rand's major research findings and final research results. The Rand Note reports other outputs of sponsored research for general distribution. Publications of The Rand Corporation do not necessarily reflect the opinions or policies of the sponsors of Rand research.

Published by The Rand Corporation

CONCEPTUALIZATION AND MEASUREMENT OF HEALTH FOR ADULTS IN THE HEALTH INSURANCE STUDY: VOL. I, MODEL OF HEALTH AND METHODOLOGY

PREPARED UNDER A GRANT FROM THE U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE

JOHN E. WARE, JR., ROBERT H. BROOK,
ALLYSON DAVIES-AVERY, KATHLEEN N. WILLIAMS,
ANITA L. STEWART, WILLIAM H. ROGERS,
CATHY A. DONALD, SHAWN A. JOHNSTON

R-1987/1-HEW
MAY 1980

Rand
SANTA MONICA, CA. 90406

PREFACE

The Rand Health Insurance Study, supported by a grant from the U.S. Department of Health, Education, and Welfare, is a social experiment being conducted in six sites across the United States to investigate the effects of different health care financing arrangements (differing coinsurance and deductible rates and fee-for-service practice versus prepaid group practice) on the use of personal medical care services, quality of care, satisfaction with care, and health status. Some 8000 people in 2750 families are enrolled in the experiment for periods of three or five years; health status is assessed for each person on entering the experiment, annually during the experiment, and on leaving.

Developing reliable and valid measures of assessing enrollee health status was a prerequisite to examination of the effects of health care financing on health status in the Health Insurance Study. The volumes that constitute Rand Report R-1987-HEW (see below) contain detailed information on the conceptualization and measurement of the health status of adults (ages 14 and older) in terms of physical, mental, and social health and general health perceptions. They also present data on the health status of adults upon enrollment in the experiment at the first site (Dayton, Ohio) and revisions made in measures of health status for repeated use in Dayton and other study sites. Measurement of physiologic health is discussed by Brook et al. in Rand's forthcoming R-2262-HEW series, which has the overall title *Conceptualization and Measurement of Physiologic Health for Adults in the Health Insurance Study.* Measurement of the health status of children (under age 14) enrolled in the experiment is discussed in a report by Marvin Eisen, Cathy A. Donald, John E. Ware, Jr., and Robert H. Brook, *Conceptualization and Measurement of Health for Children in the Health Insurance Study,* The Rand Corporation, R-2313-HEW, May 1980. Measures of adults' health habits (smoking, exercise, alcohol consumption, and weight) are discussed by Stewart et al. in the R-2374-HEW series on *Conceptualization and Measurement of Health Habits for Adults in the Health Insurance Study.*

The eight volumes in the R-1987-HEW series, which has the overall title *Conceptualization and Measurement of Health for Adults in the Health Insurance Study,* are as follows:

John E. Ware, Jr., Robert H. Brook, Allyson Davies-Avery, Kathleen N. Williams, Anita L. Stewart, William H. Rogers, Cathy A. Donald, and Shawn A. Johnston, Vol. I, *Model of Health and Methodology,* R-1987/1-HEW.

Anita L. Stewart, John E. Ware, Jr., Robert H. Brook, and Allyson Davies-Avery, Vol. II, *Physical Health in Terms of Functioning,* R-1987/2-HEW.

John E. Ware, Jr., Shawn A. Johnston, Allyson Davies-Avery, and Robert H. Brook, Vol. III, *Mental Health,* R-1987/3-HEW.

Cathy A. Donald, John E. Ware, Jr., Robert H. Brook, and Allyson Davies-Avery, Vol. IV, *Social Health,* R-1987/4-HEW.

John E. Ware, Jr., Allyson Davies-Avery, and Cathy A. Donald, Vol. V, *General Health Perceptions,* R-1987/5-HEW.

John E. Ware, Jr., Allyson Davies-Avery, and Robert H. Brook, Vol. VI, *Analysis of Relationships among Health Status Measures,* R-1987/6-HEW.

William H. Rogers, Kathleen N. Williams, and Robert H. Brook, Vol. VII, *Power Analysis of Health Status Measures,* R-1987/7-HEW.

Robert H. Brook, John E. Ware, Jr., Allyson Davies-Avery, Anita L. Stewart, Cathy A. Donald, William H. Rogers, Kathleen N. Williams, and Shawn A. Johnston, Vol. VIII, *Overview,* R-1987/8-HEW.

Volumes I-VII are directed primarily to those who will be using these measures during Health Insurance Study analyses and to other investigators who are interested in using or adapting Health Insurance Study measures for their own research. Volume VIII summarizes the results and conclusions of studies of Health Insurance Study health status measures for a more general audience. Although every attempt was made to write the volumes so that they might be read without reference to others in the series, this was not always possible. The reader is urged to consult the first volume, in particular, as it describes the model of health adopted for use in the Health Insurance Study, the site and sample selection methods, and the methods used to construct health status measures and to study their reliability and validity.

Subsequent reports will present results of revised measures of physical, mental, and social health status and general health perceptions currently in use in the Health Insurance Study.

Additional Rand reports and publications discuss other design and measurement issues related to the study. A preliminary report of issues in health status assessment appeared in Arnold I. Kisch and Paul R. Torrens, "Health Status Assessment in the Health Insurance Study," *Inquiry,* Vol. 11, 1974, pp. 40-52.

The experimental design for estimating the effects of financing on demand for care is described in Joseph P. Newhouse, "A Design for a Health Insurance Experiment," *Inquiry,* Vol. 11, 1974, pp. 5-27; and in Joseph P. Newhouse, *The Health Insurance Study: A Summary,* The Rand Corporation, R-965-OEO, March 1974. Features of the design that permit estimation of the effects on utilization behavior attributable solely to participation in the experiment are discussed in a paper by Joseph P. Newhouse, Carl N. Morris, Kent H. Marquis, Charles E. Phelps, and William H. Rogers, "Measurement Issues in the Second Generation of Social Experiments: The Health Insurance Study," *Proceedings,* Social Statistical Section, American Statistical Association, 1976.

Carl N. Morris, "A Finite Selection Model for Experimental Design of the Health Insurance Study," *Journal of Econometrics,* Vol. 11, 1979, pp. 43-61, describes the logic and techniques used to determine optimum sample sizes for the Health Insurance Study and to assign individual families to experimental plans.

The first in a projected series of reports dealing with the measurement of consumption of medical services in the Health Insurance Study is Kent H. Marquis, *The Methodology Used To Measure Health Care Consumption during the First Year of the Health Insurance Experiment,* The Rand Corporation, R-2126-HEW, August 1977. The application of reliability theory to evaluation of the quality of survey data such as those in the Health Insurance Study is discussed in M. Susan Marquis and Kent H. Marquis, *Survey Measurement Design and Evaluation Using Reliability Theory,* The Rand Corporation, R-2088-HEW, June 1977.

Other methodological issues related to techniques for obtaining precise, unbiased estimates of medical care expenditures are examined in Kent H. Marquis, M. Susan Marquis, and Joseph P. Newhouse, *The Measurement of Expenditures for Outpatient Physician and Dental Services: Methodological Findings from the Health Insurance Study,* The Rand Corporation, R-1883-HEW, April 1976.

An overview of Health Insurance Study publications is found in a paper by the same title written by Joseph P. Newhouse and Rae W. Archibald, The Rand Corporation, P-6221, December 1978.

SUMMARY

The R-1987-HEW series describes the measurement of health status of adults enrolled in Rand's Health Insurance Study (HIS). This volume presents a general introduction to the study and discusses the purpose of health status measurement. It also provides detailed information on study methods—not available in other volumes—but it does not include analytic results.

HEALTH INSURANCE STUDY DESIGN

The HIS was designed to address policy-relevant issues on the relationships among health insurance, cost-sharing (coinsurance or deductible payments), practice organization (fee-for-service or prepaid group practice), and use of health services. A sample of 7708 people in 2753 families was enrolled at six sites, which represent the four Census regions, cities of varying size, northern and southern rural areas, and locales with varying amounts of stress on the ambulatory care sector. Families were assigned to one of 16 health insurance plans that vary according to coinsurance rates; plans with nonzero coinsurance also vary according to maximum annual out-of-pocket expenditures.

Families participate for either three or five years (approximately 70 percent and 30 percent, respectively); comparison of these two subsamples will provide estimates of transitory demand. The benefits package is identical for all HIS enrollees regardless of plan and covers almost all personal health care services (ambulatory, hospital, dental, and psychological), to maximize the information gained about insuring a broad range of services.

HEALTH STATUS MEASUREMENT STRATEGY

Enrollee health status is measured in the HIS to provide data for comparative analyses of the effects of changes in the quantity and quality of personal medical care services used, across insurance plan, on the health status of a general population. Health was viewed as a multidimensional concept; physical, mental, and social dimensions of health status were identified for measurement, as was an integrative construct, that of general health perceptions. Because there was interest in understanding the effects of differences in financing on each dimension of health status, operational definitions overlapped as little as possible.

Selection of constructs within each dimension emphasized those that would be most commonly observed in adult general populations, would reflect changes in individual health that might result from changes in the use of insured services, and would be expected to show change as a function of differences in quantity and quality of services consumed during the experiment. Consistent with the World Health Organization's definition of health, HIS definitions include both positive and negative aspects of health status. Where possible, HIS measures were selected or adapted from measures that had previously been fielded in general population

studies, to allow comparison of results and to benefit from previous measurement research. Measures were evaluated against a set of criteria designed to ensure that they would provide usable data for analyses of whether and how the insurance plan affects individual health status. These criteria included such features as state-of-the-art conceptual and operational definitions, reliability, validity, and statistical power for hypothesis testing.

OPERATIONAL DEFINITIONS OF HEALTH STATUS

Operational definitions of physical, mental, and social health and of general health perceptions were developed following extensive literature reviews and consideration of the HIS measurement strategy. Physical health was defined in terms of functional status, the performance or capacity to perform a variety of activities (self-care, mobility, physical, role, household, and leisure) that are normal for individuals in good physical health. Mental health measures focused on symptoms of affective (mood) disorders and of anxiety disorders, positive well-being, and self-control, and emphasized psychological states (rather than somatic or physiological manifestations of these states). Social health was defined in terms of interpersonal interactions and activities indicating social participation. Measures of general health perceptions called for a self-rating of health in general rather than a specific dimension of health. In the HIS, these perceptions were defined with respect to time (past, present, future); perceptions of resistance to illness, worry and concern, and sickness orientation were also defined for measurement.

Volumes in the R-1987-HEW series describe the conceptualization of all health status constructs defined for measurement and include the literature reviews done to support HIS definitions and measurement evaluations. Results of evaluating the adequacy of HIS measures were based on enrollment questionnaires fielded in the first site; because these self-administered questionnaires did not include measures of social and general health, HIS results were documented only for physical and mental health status measures in the R-1987-HEW series. Measures of social and general health were included on all enrollment questionnaires after the first site and on all annual health status surveys fielded in all sites.

ANALYTIC CONSIDERATIONS

At any point in time, the physical, mental, and social health status of an individual are presumed to be interrelated, although the strength and causal nature of their relationships have not been well documented. Because of this, the HIS has presented only a simple and preliminary model of health status variables in the context of health insurance, medical care services, and other factors influencing health.

The primary focus of HIS analyses will be on the potential effect of variations in personal medical care financing on individual health status. Such effects are presumed to be a function of changes in the use and timeliness of services, the appropriateness of the provider, and the quality of the medical care process. Although many discussions assume a positive, monotonic relationship between health status and generosity of insurance, examples indicate that generous insurance may have negative as well as positive effects on health status. Consequently, hypotheses positing both effects will be tested in HIS analyses. Health status variables will be used in these analyses as dependent variables to study whether health insurance affects health status, and whether it affects various aspects of health differently. The variables will also serve as explanatory variables (covariates) in some analyses, such as those focusing on subsequent medical care consumption or patient satisfaction. Relatively simple statistical models and analytic methods will be used to the extent they are appropriate, to facilitate communication of results and their use by those involved in the health care policy process.

METHODOLOGY

This volume provides considerable detail on the methods used to select and enroll the HIS sample, collect health status data, and construct and analyze the adequacy of HIS health status measures. Eligibility criteria were broad, excluding chiefly those eligible for Medicare and institutionalized populations. Sample selection proceeded through several steps designed to determine eligibility, appropriate enrollment offers, and assignment in an unbiased manner to experimental treatment (insurance plan). Health status data are obtained primarily from six different survey instruments; all but one are self-administered, and participants receive financial incentives for completing all questionnaires. Measures of health status are obtained for each person on entering the experiment, periodically (usually annually) during the experiment, and on leaving.

Adults (ages 14 and older) who completed the Baseline Interview and the enrollment Medical History Questionnaire in Dayton provided data for constructing and analyzing HIS health status measures discussed in the R-1987-HEW series. Multi-item measures of health status variables (preferred to single-item measures) were constructed using both multitrait (modified Likert) and Guttman scaling techniques. Empirical scaling analyses were designed to make sure that the assumptions of each scaling method were satisfied by HIS health status batteries. Missing responses (which were infrequently encountered) were replaced by respondent central tendency estimates based on completed items in the same scale during multitrait analyses; visual examination of available data permitted estimates of appropriate scale scores for missing responses on Guttman scales. Evaluation of score variability identified those measures with roughly normally distributed scores. It also indicated where revisions might reduce gaps in measurement or extend the range of measurement to better detect differences in health status relevant to HIS hypothesis testing.

Three methods were used to study reliability of HIS health status measures: internal-consistency, test-retest, and reproducibility. For the measures to be useful in group comparisons, internal-consistency estimates had to meet or exceed a 0.50 standard; reproducibility coefficients (applicable only to Guttman scales) had to be at least 0.90.

Validation of HIS measures relied on several different approaches to determine whether HIS measures could be appropriately interpreted as reflecting differences in the health status dimensions they were constructed to assess. Analyses of face and content validity were used as a first step in the validation process. Because previously validated measures of the relevant constructs and agreed-upon criterion measures were not available, criterion-related validity studies were not used. In their absence, the construct validation approach provided empirical support for results from face and content validity studies. Construct validation relied on theory and empirical evidence about relationships among measures within each dimension of health and across health status measures to further understand their meaning and interpretation.

The power of HIS health status measures was evaluated to determine their ability to detect differences, if they exist, among those groups enrolled in different health insurance plans. Estimates of power for HIS physical and mental health status measures were expressed in terms of the percentage difference in mean scale scores that should be detected across plans, as well as in terms of differences in scores on health-related variables (e.g., age). These estimates were based solely on Dayton data and on certain assumptions about the distribution of the sample across plans, the method of data analysis, and conventional error rates.

ACKNOWLEDGMENTS

The authors wish to acknowledge the continuing support and cooperation of Joseph P. Newhouse, project director of the Health Insurance Study at Rand, throughout the course of the work reported in the R-1987-HEW series. In addition, we would like to thank the following individuals for their comments and suggestions during technical reviews of individual volumes:

Physical Health	Marilyn Bergner, University of Washington Sidney Katz, Michigan State University Robert L. Kane, The Rand Corporation
Mental Health	Norman Bradburn, National Opinion Research Center Harold J. Dupuy, National Center for Health Statistics Melvin Sabshin, American Psychiatric Association
Social Health	Berton Kaplan, University of North Carolina Jerome Myers, Yale University
General Health Perceptions	Marshall H. Becker, University of Michigan David Kanouse, The Rand Corporation
Relationships	Monroe Lerner, Johns Hopkins University Clairice Veit, The Rand Corporation
Power Analysis	Arthur Dempster, Harvard University

CONTENTS

FIGURES

TABLES

I. MEASURING HEALTH STATUS IN THE HEALTH INSURANCE STUDY: BACKGROUND

The eight-volume R-1987-HEW series discusses the use of survey instruments to measure health status for adults enrolled in Rand's Health Insurance Study (HIS). Volumes in the series review the literature and examine conceptual and methodological issues involved in measuring adult health status; explain the decisions made in developing the health status measures fielded in the first HIS site; present results regarding the reliability, validity, and power of these measures in testing hypotheses; and document revisions made for their subsequent use in HIS questionnaires.

This first volume provides a general introduction to the study itself and the role that health status measurement plays, and includes detailed information on study methods. The first section presents background information on the purpose and design of the HIS experiment, discusses why health status is measured in the HIS, describes the conceptual framework of health used in the HIS, and summarizes the considerations involved in selecting and defining health status variables for measurement. The second section describes the methods used to select the HIS sample, to collect health status data on adults (ages 14 and older), to construct health status measures from the first HIS health status questionnaires, and to evaluate the adequacy of these measures against criteria derived from measurement theory.[1]

DESCRIPTION OF THE HEALTH INSURANCE STUDY

The HIS was undertaken to address questions of financing health care through alternative insurance plans, and in particular to deal with policy-related issues concerning the relationships between use of health services and both cost-sharing (coinsurance or deductible payments) and practice organization (fee-for-service or prepaid group practice). Original research questions focused on microeconomic supply and demand models of health care, emphasizing measurement of how different health care financing arrangements affect the use of services. Particular attention was paid to the poor and near-poor (the disadvantaged). The initial research used nonexperimental data from health insurance carriers, national sample surveys, and other sources to develop models describing the demand for medical care services in various population groups under alternative financing or cost-sharing arrangements. Much of the nonexperimental work is now complete (see Newhouse, 1978, and Newhouse, Phelps, and Schwartz, 1974).

Available nonexperimental data on insurance and use of medical services were inadequate for the level of analysis necessary to address issues of financing health care, particularly those concerning the effects of generosity of insurance on the quality of care and health status. Consequently, a longitudinal experiment was

[1] The R-1987-HEW series discusses only the HIS measures of physical, mental, and social health and general health perceptions constructed for adults. Measures of these health dimensions developed for children (under age 14) are discussed by Eisen et al. (1980).

planned to collect more complete and accurate information. The experimental part of the HIS will provide direct estimates of the effects of differences in cost-sharing on health status, recognizing that improved health status is an important goal of efforts to expand and make more equitable the financing of medical care services. The experiment will also examine the effects of differential cost-sharing on the quality of care and patient satisfaction with care, because these variables are policy relevant and may influence the relationship between generosity of health insurance and health status.

Specific Policy Questions Addressed by the Health Insurance Study

Seven major policy research objectives were initially defined for HIS investigation. All relate to the effects of altering cost to the patient of health care on the use of services, health status, quality of care, and patient satisfaction. The objectives are enumerated below (see Newhouse, 1974, for a more detailed discussion):

1. To estimate how alternative cost-sharing arrangements affect the demand for health care services. If several groups of similar people are covered for the same health care services and the cost of these services varies from group to group (through different levels of coinsurance, deductibles, or maximum out-of-pocket payments), how does the use of or demand for health services vary across those groups?
2. To assess the effect of varying the out-of-pocket cost of health services on individual health status. If the use of health services differs as a function of cost, what is the effect on health?
3. To determine whether and by how much cost-sharing arrangements affect low-income families more than higher-income families.
4. To learn whether the quality of the medical care process differs for individuals with various health insurance financing plans.
5. To ascertain how the ambulatory care system responds to varying levels of demand or stress. Differences in the use of services, quality of care, and patient satisfaction can be examined as a function of such factors as delays in making appointments, waiting time in physicians' offices, or referral patterns.
6. To compare utilization, health status outcomes, quality of care, and patient satisfaction in an existing prepaid group practice and in the fee-for-service system.
7. To gain familiarity with the difficulties of administering health insurance plans that relate the degree of cost-sharing to the patient's income.

Health Insurance Study Design

Sites. To accomplish the HIS objectives, a sample of 7708 people in 2760 families were enrolled at six sites across the country: Dayton, Ohio; Seattle, Washington; Fitchburg, Massachusetts; Franklin County, Massachusetts; Charleston, South Carolina; and Georgetown County, South Carolina. Table 1 gives the number of families and individuals originally enrolled in each site and the enrollment dates. Exit dates follow three or five years later, depending on the enrollment period.

Table 1

NUMBER OF INDIVIDUALS AND FAMILIES ENROLLED IN THE HEALTH INSURANCE STUDY BY SITE, ENROLLMENT DATE, AND ENROLLMENT PERIOD

		Enrollment Period			
		Three Years		Five Years	
Site	Enrollment Dates	Individuals	Families	Individuals	Families
Dayton	11/74-2/75	538	186	602	204
Seattle					
Fee-for-service	1/76-9/76	919	359	301	125
Prepaid group practice	4/76-9/76				
Experimentals		579	229	562	219
Controls		--	--	751	304
Massachusetts	6/76-10/76				
Fitchburg		548	191	176	58
Franklin Co.		651	241	240	76
South Carolina					
Charleston	11/76-1/77[a]	572	194	208	68
Georgetown Co.	11/78-1/79[b]	798	231	263	75

[a] Enrollment dates for five-year group at both South Carolina sites.

[b] Enrollment dates for three-year group at both South Carolina sites.

Sites were selected to represent the four Census regions of the country, cities of varying sizes, both northern and southern rural areas, and locales with varying amounts of stress on the ambulatory care sector. (At some sites, delays for new and/or return appointments were long, while at others they were trivial or nonexistent.) Other site selection criteria were that each metropolitan area was to be within one state; one site was to have a long-standing prepaid group practice willing to participate in the HIS (i.e., allow people to be randomly assigned to it as one of the experimental plans); and, all other things being equal, preference was given to areas with lower medical prices. For cities, the urbanized area within the Standard Metropolitan Statistical Area (as defined by the Bureau of the Census) was the site of interest; hence, Seattle includes the Seattle-Everett urbanized area and Fitchburg includes Fitchburg-Leominster.

Experimental Treatments. The 16 health insurance plans that constitute HIS experimental treatments vary according to levels of cost-sharing (coinsurance rates); plans with nonzero coinsurance rates also vary according to maximum out-of-pocket expenditures in any one year. (After the maximum is met in these plans, the coinsurance rate becomes zero for covered services.) The experimental treatments include

- One plan in which care is free to the family.
- Three plans with 25 percent coinsurance (i.e., the family pays 25 percent of its medical bills).
- Three plans with 50 percent coinsurance (two of these only in Dayton).
- Three plans with 50 percent coinsurance for dental and outpatient mental services and 25 percent coinsurance for all other services (all sites except Dayton).

- Three plans with 95 percent coinsurance (100 percent in Dayton during the first year of the experiment).
- One plan with 95 percent coinsurance (100 percent in Dayton during the first year) up to a maximum expenditure of $150 per individual, or $450 per family, each year and no coinsurance above that. In this plan only, the coinsurance applies solely to ambulatory expenditures; inpatient expenditures are not subject to coinsurance.
- One plan that assigns some of the Seattle participants to a prepaid group practice (Group Health Cooperative of Puget Sound) in that site.
- One plan (a control group) that is a sample of people who met HIS eligibility criteria and were already enrolled in the Seattle prepaid group practice (to study whether those who have self-selected a prepaid group practice systematically differ from those who have not).

All plans except the first and last three have an income-related ceiling on annual out-of-pocket expenditures paid by the family. This "maximum dollar expenditure" (MDE) is 5, 10, or 15 percent of annual family income up to a limit of $1000 per family per year for the 50 and 95 percent coinsurance plans and $750 for the 25 percent plans (the maximum was $1000 in Dayton for the first two years and in Seattle for the first year). Because the MDE is directly proportional to family income, income was carefully and comprehensively defined (see Clasquin and Brown, 1977).

Table 2 reports the number of individuals and families enrolled by plan as of the end of the enrollment process in each site. Table 3 gives the distribution of individuals and families by plan.

Families were enrolled in one of the HIS insurance plans for either three or five years (approximately 70 percent and 30 percent of the sample, respectively). Comparison of the two samples will provide information about the extent of transitory demand. In each site except South Carolina, participants were enrolled at the same time, and exit dates for the two samples were set two years apart; hence, transitory demand will reflect services either "crowded into" the last year or postponed until after participation ends. For South Carolina, the enrollment dates differed by two years and the exit dates were the same; thus, comparison of the two groups from this site will reflect initial "catch-up" demand. In short, data on the five-year sample will permit estimates of transitory demand in the three-year sample, thus allowing use of the data from the entire third year of the experiment (appropriately adjusted), even though a large part of the total sample is leaving.

When families were enrolled, they agreed to assign the benefits from any existing health insurance policies to the HIS during their enrollment and to allow the HIS to claim reimbursement from those policies for services covered in common. The assignment included benefits from policies that might be acquired subsequently. These policies were held in escrow and premiums were paid until the end of the enrollment period. If a family's former insurance coverage was more generous than that of the HIS plan, the family received monthly payments to ensure that it would never be financially worse off under the HIS plan. Certain payments were also made to families for filing regular reports on health care use and for completing HIS interviews. These and other operational rules are explained in detail by Clasquin and Brown (1977).

Table 2

Number of Individuals and Families Enrolled in the Health Insurance Study by Insurance Plan Characteristics and Site

	Sites											
	Dayton, Ohio		Seattle, Wash.		Fitchburg, Mass.		Franklin Co., Mass.		Charleston, S.C.		Georgetown Co., S.C.	
Insurance Plan[a]	Individuals	Families	Individuals	Families	Individuals	Families	Individuals	Families	Individuals	Families	Individuals	Families
0% coinsurance (no MDE)	301	95	430	162	242	78	297	104	264	92	358	105
25% coinsurance												
5% MDE	95	32	58	22	9	3	19	8	26	7	23	6
10% MDE	86	27	38	17	17	5	25	10	15	6	44	11
15% MDE	79	28	36	16	11	4	17	7	27	8	21	8
50% coinsurance												
5% MDE	65	27	--	--	--	--	--	--	--	--	--	--
10% MDE	62	18	--	--	--	--	--	--	--	--	--	--
15% MDE	67	26	--	--	56	17	58	19	26	9	52	15
25%/50% coinsurance												
5% MDE	--	--	44	19	27	9	25	8	25	10	43	14
10% MDE	--	--	49	20	31	10	35	14	34	12	41	11
15% MDE	--	--	28	13	30	10	31	7	20	7	28	7
95% coinsurance												
5% MDE	107	34	89	33	38	14	60	19	39	14	47	13
10% MDE	80	30	69	28	40	16	50	19	40	12	46	14
15% MDE	93	34	94	35	35	14	54	20	68	19	76	21
95% coinsurance ($150/$450 MDE)	105	39	285	119	188	69	220	82	196	66	282	81
Total fee-for-service	1140	390	1220	484	724	249	891	317	780	262	1061	306
Prepaid group practice	--	--	1892	752	--	--	--	--	--	--	--	--
Overall total	1140	390	3112	1236	724	249	891	317	780	262	1061	306

[a] MDE = maximum dollar expenditure; see text for explanation.

Table 3

TOTAL NUMBER AND PERCENTAGE OF INDIVIDUALS AND FAMILIES ENROLLED IN THE HEALTH INSURANCE STUDY BY INSURANCE PLAN

	Individuals			Families		
Insurance Plan[a]	Number	Percentage of Fee-for-Service Sample	Percentage of Total Sample	Number	Percentage of Fee-for-Service Sample	Percentage of Total Sample
0% coinsurance (no MDE)	1892	32.5	24.5	636	31.7	23.0
25%, 50%, and 25%/50% coinsurance (5%, 10%, 15% MDEs)	1523	26.2	19.8	527	26.2	19.1
95% coinsurance (5%, 10%, 15% MDEs)	1125	19.3	14.6	389	19.4	14.1
95% coinsurance ($150/$450 MDE)	1276	21.9	16.6	456	22.7	16.5
Total fee-for-service	5816	99.9[b]	75.5	2008	100.0	72.8[b]
Prepaid group practice[c]						
Experimentals	1141	--	14.8	448	--	16.2
Controls	751	--	9.7	304	--	11.0
Total sample	7708	--	100.0	2760	--	100.0

[a] MDE = maximum dollar expenditure; see text for explanation.

[b] Percentages may not total because of rounding.

[c] Seattle only.

Benefits. The package of benefits is identical for all HIS participants regardless of plan. It is extremely comprehensive, and covers ambulatory and hospital care, nursing home care, preventive services, all dental services except orthodontia, prescription drugs and appliances, certain over-the-counter drugs when prescribed by a physician, psychiatric and psychological services, and almost all other personal medical services, including many that are not typically covered by conventional health insurance plans (such as care provided by chiropractors and Christian Science healers). In general, the benefits package was designed so the HIS could maximize the information gained about the effect of insuring many different services. The primary exception to this rule was a service, such as orthodontia, for which a large amount of transitory demand could be anticipated.

CONCEPTUAL FRAMEWORK AND STRATEGY FOR MEASURING HEALTH STATUS

In the HIS, health is viewed as a multidimensional concept. Following the definition of health proposed by the World Health Organization (1948)—that "health is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity"—three of the dimensions identified for measurement were physical, mental, and social health. An integrative concept not specified by the WHO, general health perceptions, was also included among HIS health status measures because it was believed to reflect all health dimensions and to contain unique subjective information about health. In addition to measuring health status per se, several health-related constructs were also selected for measurement because they are relevant to HIS hypotheses. Among these are several variables referred to as "health habits": in particular, the HIS obtains self-reports of behaviors related to smoking, overweight, alcohol consumption, exercise, and safety practices (Stewart, Brook, and Kane, 1979; Stewart, Kane, and Brook, 1980; Kane, Brook, and Kane, forthcoming; Stewart, Kane, and Brook, forthcoming).

One aspect of physical health—that of physiologic health (the status and functioning of specific organ systems)—was singled out for separate measurement. Measuring physiologic health was considered important because progressive disease of organ systems can occur without immediately producing overt symptoms or changes in behavioral manifestations of functional status and may not be detected by measures of physical, mental, and social health. These diseases may and usually do lead to decreases in one or more of these three dimensions of health status. Many are treatable or curable with high-quality medical care, which may in turn influence their effect on other dimensions of health (e.g., functional abilities or mental and social adaptation). Brook and his colleagues discuss the operational definition of physiologic health in the HIS in a companion series (R-2262-HEW, forthcoming). That series presents the rationale for measuring the presence and severity of some 25 specific diseases and conditions as part of the overall HIS health status measurement strategy. The actual procedures and laboratory tests done as part of the disease-specific assessments are described in the screening examination manual (R-2101-HEW) by Smith et al. (1978).

Before a more detailed discussion of HIS health status concepts, a review of the goals of measurement and constraints on data collection in the HIS is pertinent.

These factors influenced HIS selections among the many available operational definitions of the health status dimensions and should be kept in mind when evaluating HIS selections and definitions.

HIS Measurement Strategy

Several features of the HIS study design and the planned use of health status measures had implications for the health status measurement strategy. HIS health status measures were developed chiefly to provide data for comparative analyses of the effects of changes in the quantity and quality of personal medical care services, across health insurance plans, on the health status of a general population. In view of the interest in analyzing the effects of differences in financing on specific health status dimensions, the HIS focused on measuring each one as separately as possible, minimizing overlap in operational definitions of each major dimension and of each construct within a given dimension. The operational definitions of physical, mental, and social health status dimensions that were deemed appropriate for HIS use were, therefore, not as comprehensive as others found in the literature, and they will not prove useful for all research purposes.

In selecting constructs within each dimension for measurement, emphasis was placed on those that would be most commonly observed in adult general populations, and that would be expected to show change as a function of differences in the quantity and quality of medical services consumed during the experiment. Many health problems are common in adult general populations. Analyses of the effects of differences in health care financing on health status required that the constructs selected for measurement reflect changes in an individual's health that might result from varying the use of services covered by health insurance plans that differ in the amount of out-of-pocket expenditures required. Although the health and well-being of individuals is influenced by community maintenance activities, public health services, and the general performance of educational and other social systems, such activities are not within the scope of the personal medical care delivery system. Furthermore, they are not covered by health insurance. Thus, measures that may indicate community health (e.g., population density, housing conditions) were not included in the HIS measurement strategy.

The HIS was not limited to using health status measures that would yield one number representing an adult's overall health status. Batteries of items corresponding to the three major dimensions of health status, physiologic health, and general health perceptions are each scored and interpreted separately. Within each battery, items are included to measure more than one construct (e.g., within the battery used to measure mental health, items are included to measure anxiety, depression, positive well-being, and self-control). If analyses indicate that the information provided by several measures of health status (within each battery or across batteries) can be summarized in a smaller number of scores without significant loss of information, such a composite or aggregate measure may be constructed post hoc.

Following WHO guidelines and contemporary definitions, both negative and positive aspects of health were assessed whenever possible. Medical care as currently practiced may be more likely to affect the negative end of each spectrum. Emerging public interest in positive health and holistic medicine (reflected in the comprehensive set of benefits covered by the HIS) suggested that efforts also be directed to measuring more positive aspects of health. No attempt was made to measure extremely positive aspects of health that might be affected by services obtained by only a few people in the study, or in society, or that could be achieved by a relatively small number of people.

Because longitudinal data on health status are available in the HIS (collected at enrollment, on biweekly health reports, on annual health questionnaires, and at exit), development of physical, mental, social, and general health status measures that explicitly included prognosis was considered unnecessary. Moreover, the predictive ability of these HIS health status measures (beyond the termination of the study) could be enhanced by using information from physiologic health and health habits measures. For example, epidemiologic evidence supports the view that the current practice of certain negative or positive health habits (e.g., smoking, maintaining normal weight) is related to future states of worse or better health. Thus, measuring individual health habits in the HIS effectively extends the length of the experiment and makes it more powerful, because changes in health habits might be observed in the experimental period even if related shifts in health status (particularly physical health) were not.

Constraints on data collection (e.g., the potentially sensitive nature of some questions about health) indicated that mailout and mailed return of self-administered questionnaires were preferred. Self-administration enhances privacy because each adult (14 years old and older) completes the Medical History Questionnaires (MHQs) and the Health Questionnaires in private and seals them in an envelope. This practice helps to keep each person's responses from being seen by others in the family.[2]

The self-administration strategy was adopted when analyses indicated that it would provide data of acceptably high quality, when combined with assistance and follow-up if needed, even among the least educated groups enrolled in the HIS. Stringent edit specifications are used while the questionnaires are still in the field. For example, respondents are contacted if the response to a key question (e.g., the initial question in batteries with skip patterns) or responses to six or more questionnaire items are missing. These specifications identify gross problems in data quality (i.e., missing responses) while there is still an opportunity to correct them.

Moreover, reliance on self-report for the annual health status surveys makes it possible to gather information about many aspects of health from all HIS participants, rather than the smaller subset that would necessarily have been studied had personal interviews or clinical examinations provided all health status data. In addition, the lower costs associated with self-administered instruments permitted enrollment of a larger sample for a given budget. It was also assumed that participants might respond more frankly to sensitive questions about personal health status if questionnaires were self-administered rather than interviewer-administered.

Furthermore, the HIS has not encountered the typical problems of high rates of questionnaire nonresponse often associated with self-administration. Completion of all questionnaires is a condition of enrollment, and respondents are reimbursed small amounts for filing completed questionnaires (see Sec. II for further details).

[2] Privacy and confidentiality of information on individuals and families in the HIS data base as a whole are preserved through an elaborate three-part numbering system, only two parts of which are known to any person working with the data.

These conditions and methods have produced very high response rates—generally 90 percent or above.

Finally, the HIS required, to the extent possible, that the health status measures be selected from those previously fielded in studies of general populations or be adapted from such measures. This requirement allows comparison of HIS health status measures with those used in national probability sample surveys, such as the National Health Interview Survey, and enables the HIS to benefit as much as possible from previous measurement research.

To ensure that HIS measures of health status would successfully fulfill their intended purpose—providing data to allow analysis of the effects of different health insurance plans on individual health status—the measures had to meet the following criteria:

- The measures should agree with contemporary conceptualizations of the three major dimensions of health and of constructs within those dimensions.
- The operational definitions of each dimension and construct represented by HIS questionnaire items should reflect the state of the art of measurement as defined in the literature.
- The items used to measure each construct should be combined in such a way that the number of variables or scores used to define that health status construct is reduced as much as possible without substantial loss of information.
- Score distributions for each measure should have sufficient variability to be useful in detecting actual differences in health status of people in a general population for whom repeated measures are available (i.e., should have sufficient power to test hypotheses about differences in health status as a function of differences in health insurance plans).
- The measures should be substantially free of error (i.e., be as reliable as necessary) to allow confident estimation of average levels of health status within groups and comparison between different groups in the enrolled population (e.g., between plans or between disadvantaged and nondisadvantaged groups).
- Each measure should provide information about the particular health dimension or construct it was intended to measure (i.e., be valid) without duplicating information obtained from other HIS health status measures.

These criteria also had to be fulfilled in subgroups of the HIS sample for which data quality could be expected to be poorest, such as enrollees who were disadvantaged in terms of education and income.

Operational Definitions of Major Health Dimensions

This section summarizes HIS definitions of the three major health status dimensions (physical, mental, and social) and of general health perceptions. These definitions were adopted after considering the WHO recommendations for conceptualizing health status and the HIS measurement strategy outlined above.

Physical Health. In HIS health questionnaires, physical health has been operationally defined in terms of functional status.[8] Functioning refers to performance of or capacity to perform a variety of activities that are normal for an individual in good health. A review of the literature (see Vol. II) on measures of functional status identified six categories of activities for which performance or capacity has been assumed to reflect primarily a person's physical, rather than mental or social, health. These include self-care activities (e.g., feeding, bathing); mobility (e.g., confinement indoors); physical activities (e.g., walking, running); role activities (those typical for an individual of a specified age and social role such as job or school); household activities; and leisure activities (e.g., hobbies, clubs). Measures of performance and/or capacity in all six categories have been included in HIS batteries of items hypothesized to measure physical health. Published studies suggest that activities included in the role and leisure activity categories may overlap conceptually with those thought to reflect mainly social health (i.e., social participation and interpersonal interaction) (see Vol. IV). These activities were included, however, as hypothesized measures of physical health in the first HIS site (Dayton) and whether they measured physical or social health was explored empirically.

Mental Health. The HIS strategy for measuring mental health emphasized assessing phenomena of psychological disorders about which there was considerable conceptual agreement, which occurred commonly in general populations, which might be responsive to changes in the quality or quantity of mental health services consumed during a three- or five-year period, and which could be quantified from self-administered questionnaires. Because of these constraints, many mental health constructs (e.g., schizophrenia) that have received considerable attention in the theoretical and empirical literature (see Vol. III) were not candidates for measurement. Thus, the operational definition of mental health adopted for the HIS was not as comprehensive as might be wished, particularly if mental health were the sole dependent variable measured in the experiment.

In view of these constraints, importance was attached to measuring symptoms of affective (mood) disorders, such as feelings of depression, and of anxiety disorders. Consideration was also given to assessing positive aspects of mental health that have received increasing attention in the literature, including positive well-being and, to a lesser extent, self-control of behavior, mood, thoughts, and feelings. Operational definitions of these mental health constructs in the HIS focused chiefly on psychological states, rather than on physiological and somatic states (i.e., physical manifestations of mental problems), and on both favorable and unfavorable aspects of these states. As noted above, constraints of the measurement strategy precluded development of measures that would reflect etiology, specific diagnostic information, or prognosis.

Social Health. Social health has been viewed in the literature (see Vol. IV) both as a distinct dimension of health status (i.e., as a dependent variable) and in terms of social support systems that modify the effect of the environment and stressful life events on physical and mental health (i.e., as an intervening variable). The literature on conceptualization of social health indicated less consensus than

[8] Physical health has also been operationally defined in terms of disability days on biweekly diaries. The physiologic component of physical health, which was singled out for separate measurement, is defined in terms of the absence or presence and severity of some 20 chronic diseases; data come from the MHQ and the results of a medical screening examination (see Brook et al., forthcoming). Only the survey measures of functional status are discussed in the R-1987-HEW series.

the literature on physical health as to categories of activities that reflect primarily an individual's social health. There appeared to be some agreement, however, that social health could be operationally defined in terms of interpersonal interactions (e.g., visits with friends) and activities indicating social participation (e.g., membership in clubs). The measures include both objective reports (e.g., counts of number of friends and memberships) and subjective ratings (e.g., how well one is getting along with others). Social health appears to differ from physical and mental health constructs because it extends the definition of health beyond the physiologic, physical, and psychological status of the individual and focuses on the quantity and quality of interpersonal ties and extent of community involvement.

Because identifying an appropriate measure of social health was difficult, a battery of items hypothesized to reflect primarily social health was not included in the Dayton enrollment MHQ. (That questionnaire did include items that may be sensitive to social health differences in behavioral terms in the role activities category of HIS physical health measures and in emotional terms in the love relationships category of HIS mental health measures.) A battery designed specifically to measure social health was included in all HIS health status questionnaires fielded after the Dayton enrollment MHQ, and studies of its measurement properties will be documented in a subsequent Rand report.

General Health Perceptions. HIS measures of general health perceptions differ from other measures of health status in that they do not focus on a specific dimension of health status (i.e., physical, physiologic, mental, or social). Instead, such measures ask respondents for an assessment or self-rating of their health in general. In theory, this difference in measurement makes it possible both to assess the objective information people have about their health and to take into account the cognitive process underlying an evaluation of that information. In the HIS, general health perceptions have been defined with respect to time (perceptions of prior, current, and future health) and with respect to three other constructs indicative of general health perceptions, including resistance to illness, health worry and concern, and sickness orientation (the extent to which people perceive illness to be a part of their lives). Both favorable and unfavorable definitions of health were included in the operational definitions of these constructs.

A comprehensive battery of general health perception items was not fielded in the Dayton enrollment MHQ; three items in that questionnaire focused on a general evaluation of current health (as excellent, good, fair, or poor), worry/concern, and pain. A full battery of items designed specifically to assess several aspects of general health perceptions was included on all HIS health questionnaires fielded following the Dayton enrollment MHQ, and measurement studies will be documented in a subsequent Rand report.

ANALYTIC FRAMEWORK FOR HEALTH STATUS VARIABLES

Overview

Although physical, mental, and social health were conceptualized as distinct dimensions of health status, they clearly interrelate to some largely unknown degree. Many descriptions of their relationships have been offered, as illustrated in the following examples. An individual with acne may withdraw from social contact, which may in turn lead to depression or anxiety. A person's loss of a limb will decrease his physical abilities, which may subsequently result in decreased participation in social activities and more emotional stress and unhappiness. A person experiencing emotional distress may be more likely to experience physical injury. Thus, the best overall model of health seems to be one that specifies an underlying general health concept, possibly that defined by general health perceptions, common to the physical, mental, and social dimensions of health status. This model should also specify that each of these dimensions can be evaluated separately despite their interrelationships.

Figure 1 presents a Venn diagram of three overlapping circles that illustrates how health status has been viewed in the HIS. This model is preliminary and is only partially specified; for example, the strengths of relationships among components are not quantified. Very little has been published regarding how health status dimensions relate to each other (see Vol. VI); with rare exceptions, comprehensive batteries including measures of the three major health status dimensions have not been fielded in general populations. The formulation shown in Fig. 1 relied on theory and on comparisons of results across published studies differing in methods and population characteristics.

As implied by the hatched area in Fig. 1, the HIS model includes a general health status factor that accounts for variance in all three major health status dimensions. Personal assessments of general health status (e.g., the frequently studied rating of current health as excellent, good, fair, or poor) were hypothesized to assess primarily this general health factor. Tests of this hypothesis were not found in the published literature. However, cross-sectional evidence supporting the overlap between each pair of circles in Fig. 1 has been published. For example, physical and mental health have been positively correlated (Bradburn, 1969; Gilson

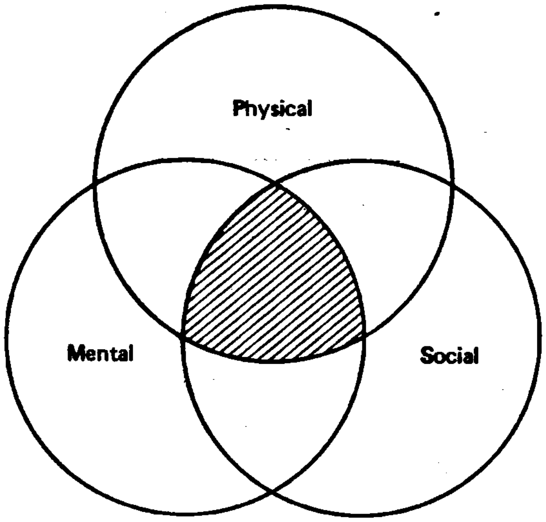

Fig. 1—Hypothesized relationships among three major dimensions of individual health status

et al., 1975; Social Psychiatry Research Unit, 1977); physical health has been linked to social health (Jeffers and Nichols, 1961; Palmore and Luikart, 1972; Fine, 1975; Greenblatt, 1975); and mental and social health have been positively correlated (Bradburn, 1969; Klemmack, Carlson, and Edwards, 1974; Fine, 1975; Greenblatt, 1975).

Figure 1 has been drawn so that the three circles corresponding to physical, mental, and social dimensions of health status overlap equally with each other. This is probably not the case. As discussed above, the hypothesis that all three dimensions (favorably defined) are directly related to each other at a point in time is supported by the published literature. For methodological reasons (discussed in Vol. VI), it is not possible to infer from these studies the actual strength of these interrelationships (i.e., the amount of overlap in the circles in Fig. 1) and to clarify the nature and extent of the causal relationships involved. Results documented in Vol. VI suggest that objectively defined social health variables tend to vary more independently than physical and mental health scores (regardless of how the latter are defined), and that social health is the least related of the three dimensions to general health perceptions. These results suggest that Fig. 1 might be more accurately drawn with the circle defining social health overlapping less with the other two circles.

The relationship of the physiologic dimension to the three major health status dimensions is twofold. In some cases, such as with very poor lung or heart function, clinical disease (e.g., emphysema, heart failure) is present and may be irreversible. Although the disease may be adequately managed and its effect on the patient's daily life minimized, its presence is likely to affect the person's physical, mental, and social health directly. In other cases, such as subclinical renal dysfunction or borderline low hematocrit, no readily identifiable disease is present or known to the patient; however, when stressful events occur, they may be handled less well than if a subclinical problem were not present. In such cases, physiologic health status measures may tap "precursor" variables that come into play only over time with aging or with stress.

Figure 2 presents a very simple depiction of HIS analyses related to health status; it puts the health status model into the context of the medical care process and other factors affecting health. In addition to their effects on each other (Fig. 1), the dimensions of health status are affected, separately or in combination, by another set of factors—environmental conditions; genetics, including biologic variability among organisms; public health level of the community; and personal behavior and life-style, including health habits. Many variables defining these factors could be hypothesized to affect health status, but with the possible exception of life-style, they fall outside the domain of personal medical services. Although their importance to national health policy should not be minimized, the policy issues addressed by the HIS are confined largely to variations in health as a direct or indirect function of variations in financing of personal medical care. Changes in the way personal medical care is purchased (by manipulations through health insurance of the price to be paid by the patient) are unlikely to produce changes in such factors as community environment or heredity.

Because the HIS emphasizes the potential effect on an individual's health status of variations in the level of personal medical care financing, the analytic model must specify ways in which such an effect might occur. As shown in Fig. 2,

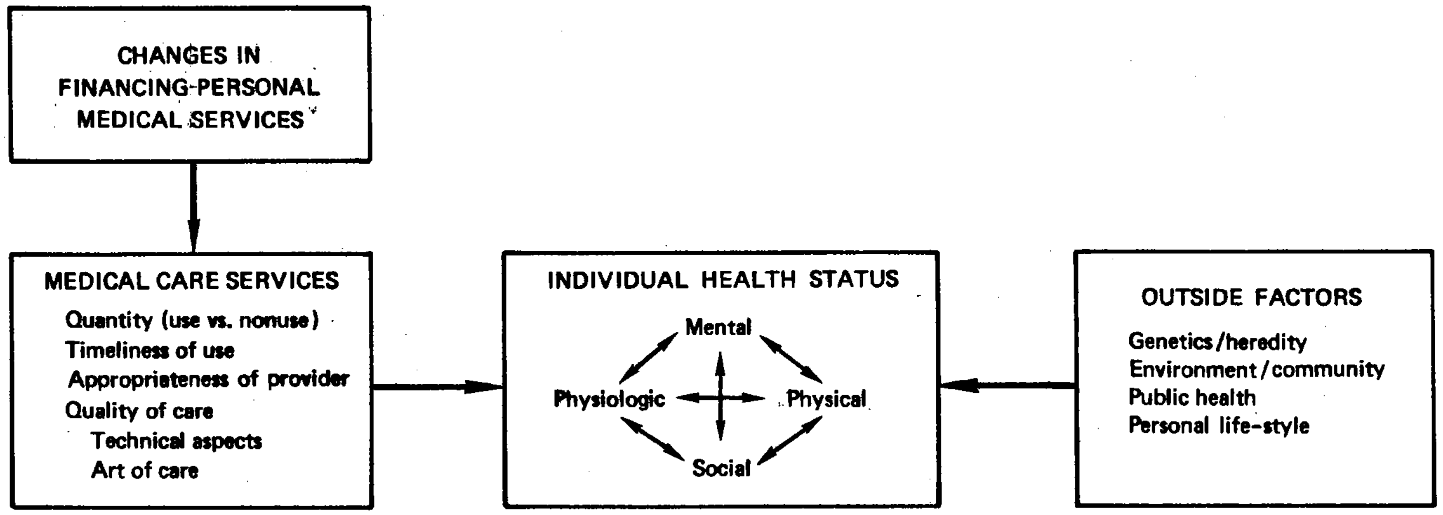

Fig. 2—Preliminary framework for analyzing the effects of health care financing on individual health status

one or more of the following variables must change to produce improvements in health status: quantity of services obtained; timeliness of services; appropriateness of provider; and quality of services (technical level and art of care).

The following example may clarify the relationship between these variables and health status. Consider an adult with a productive cough, fever, and probable bacterial pneumonia. If the family is on the zero coinsurance plan, medical care has no financial cost to the family. The adult may therefore seek medical attention when under other circumstances he might not have done so or might have delayed seeking care. If medical care is obtained during the illness episode, health status (e.g., measured by avoidance or minimization of pulmonary complications) may improve. By implication, if the person receives timely medical attention, that is, soon after symptoms begin rather than after lung tissue or lining is destroyed, his eventual health status is more likely to be better. Moreover, if the person is treated by an appropriate provider (a physician trained to treat the particular problem or the particular age group, which in this example is an internist rather than a surgeon), better care should result, also leading to better health status outcomes. Finally, if the quality of care received is better under generous health insurance, all other things being equal, health status outcomes are likely to be higher. For instance, if the technical quality of care is higher (e.g., the appropriate antibiotic is given in the correct dose and for an adequate period of time) and if the art of care is higher (e.g., physician sensitivity to patient concerns leads to greater compliance with the medications regimen), health status outcomes should be better.

Discussions of the relationship between health status and generosity of health insurance often implicitly assume that such a relationship is positive, although not necessarily linear. This is probably an overly naive view. More generous health insurance may be a two-edged sword. On one hand, more generous insurance by definition improves financial aspects of access and may also improve nonfinancial aspects of access, such as enabling the patient to choose a provider whose office is more convenient, or to make an appointment rather than wait in an emergency room. Both effects would be expected to increase the use of services, which, if necessary and appropriate, may in turn improve health status. Moreover, as changes in access and utilization are observed, quality may improve when physicians, uninhibited by their patients' financial constraints, are able to order necessary diagnostic procedures and provide more comprehensive care and follow-up. These effects may also result in improved health status.

On the other hand, the same effects—increased access, use, and intensity of services—may adversely affect health status. Increased access, for example, may prompt overuse of services and inappropriate sick role behavior; greater use and intensity of services may result in improper labeling of persons as ill and in iatrogenic disease. All such consequences are threats to improved health status that may result from more generous health insurance coverage. Plausible hypotheses about the relationship between health status and generosity of insurance, therefore, might include those positing a negative relationship, as well as the more commonly hypothesized positive relationship.

An example may clarify these points. Consider the person who suffers somatic symptoms and fatigue because of a stressful life event. With free care (the zero coinsurance plan in the HIS), the person might choose to talk to a physician, especially in the absence of a network of close friends. If care involved financial cost (a plan with coinsurance), the person might not seek medical advice or care. In the former case, the doctor might be interested primarily in detecting physical illness and might perform a series of diagnostic tests. In all likelihood, the results would be equivocal, in that physical illness may not be ruled out completely, which in turn might lead the physician to legitimize the patient's adoption of the sick role. Adoption of the sick role might be observed in such behaviors as taking time off from work or other daily activities, or in increased worry or concern about having an unknown disease. If these hypotheses were borne out by data collected in the HIS, they might indicate a negative relationship between generous health insurance and health status.

As noted above, measurement of variables that are completely exogenous to the personal medical care system (e.g., public health services, community health) was not considered germane to HIS analyses. All other variables identified in the preceding analytic framework are measured, and the resulting data allow hypotheses about the relationship of medical care financing and health status, as moderated by quantity, timeliness, appropriateness, and quality of care, to be tested.

Analytic Considerations

In HIS analyses, the health status variables will serve two functions. As depicted in Fig. 2, they will be used as dependent variables to address the question: Does health insurance have different effects on the various dimensions of health? For example, does the zero coinsurance plan have its greatest influence on mental health because families on the plans with coinsurance are reluctant to incur the costs of mental health services but view other medical services as necessities?

The health status variables will also be used as explanatory variables (covariates) in some analyses, such as economic analyses that focus on medical care consumption, analyses of health-related variables such as patient satisfaction or health habits, and analyses of other dimensions of health status. For instance, the presence of certain chronic diseases that impair physical functioning, such as arthritis, may require statistical control when studying differences among plans in level and type of services used, differences among subsamples in satisfaction with care, or differential shifts in mental health.

The statistical analyses performed to test HIS hypotheses will remain as simple as possible to facilitate their communication and use by those involved in the health care policy process. For example, to study the effects of differences in health care financing on health status, analyses will begin with a simple multiple linear regression of one health status variable measured at the end of the experiment on generosity of health insurance, other experimental variables (e.g., whether a screening examination was administered at enrollment), demographic and socioeconomic variables, and the initial measure of the health status variable taken at enrollment.

Many factors, such as interactions among variables, may necessitate a more complex analytic strategy. For instance, the effect of generous health insurance on medical care consumption and health status may depend on whether enrollees believe that medical care providers exercise substantial control over health and illness. Thus, the analytic model may require the addition of terms defining interactions between insurance plan and the personal beliefs of enrollees to fully understand plan effects. These and other factors will be evaluated in detail before relying solely on simple models and analytic methods.

To improve measurement validity and to reduce the number of required analyses, questionnaire items were aggregated into multi-item scales that define the major constructs within the physical, mental, and social dimensions of health status. To permit a focused analysis, hypotheses will be tested using a single measure of each construct as the dependent variable before constructs are further aggregated within dimensions. For example, with regard to mental health services, the effects of health care financing and organization on distinct mental health constructs (e.g., anxiety, depression, and positive well-being) will be tested before using an overall mental health index to summarize the results of mental health analyses. If these analyses indicate that conclusions about the effects of health care financing on mental health do not differ by the mental health construct, a summary mental health index will be used.

II. METHODOLOGY: SAMPLING, DATA COLLECTION, AND CONSTRUCTION AND ANALYSIS OF HEALTH STATUS MEASURES

This section includes detailed descriptions of the methods used in the HIS to (1) identify, select, and enroll experimental subjects; (2) obtain health status data; (3) construct multi-item health status scales; and (4) evaluate the adequacy of scales in terms of reliability, validity, and power for purposes of testing HIS hypotheses. The detail presented here supports the brief discussion of these methods in other volumes in the R-1987-HEW series.

ELIGIBILITY CRITERIA AND SAMPLING PROCEDURES

The individuals offered enrollment in the HIS in Dayton were not a random sample of their metropolitan area, because not all people were eligible. There were several ineligible groups: families with household heads 62 years old and older (60 years in the case of families enrolled for five years) and persons who qualified either for Medicare or for Supplemental Security Income (SSI) after its inception. The first group was ineligible because family heads would qualify for Medicare because of age before the end of the experiment. The Medicare population was not included because of the different nature of its health problems and the existence of the Medicare program. The SSI population was excluded because it was not possible to obtain a waiver of HIS monies paid to SSI recipients, and HIS benefits would have had to be deducted from their SSI grants; thus, participation might have created a financial burden for these people.

Also ineligible were persons who qualified for military health care or certain public sector benefits under the Indian Health Service. Many such persons could obtain medical services outside the HIS (which would confound the analyses) and would probably continue to do so were a national health insurance plan enacted. Those institutionalized for an indefinite period, such as prisoners or patients in state mental institutions, were also ineligible.

The Medicaid population was eligible. Those persons 62 and over (in the three-year sample) or 60 and over (in the five-year sample) who were members of a household with an eligible head were enrolled in the study for data collection purposes, but they did not receive insurance benefits.

Selection of the sample to be offered enrollment, therefore, included several steps designed to determine eligibility. First, a sample of families completed a short screening interview, giving information pertinent to eligibility. From those determined to be eligible, a stratified random sample was drawn. This group received a Baseline Interview to determine who could be offered enrollment and the type of enrollment offer that could be made, and to obtain other information necessary to assign families to experimental treatments (plans) according to the design. During enrollment, some families selected could not be located, others were no longer eligible, and some refused the enrollment offer. Thus, the composition of the final sample—those actually enrolled in Dayton—differed somewhat from the group receiving the Baseline Interview; these differences were not statistically significant. These steps, the numbers of people involved, and the characteristics of the final Dayton sample are described in greater detail below.

Screening Interview Sample

The urbanized area of Dayton, Ohio (as designated by the Census Bureau) was divided into 15 geographic regions of approximately equal population, based on information from the 1970 Census. Each of these 15 regions or strata was subdivided into about 150 clusters of 300 people each. Fifty-seven clusters were then randomly selected (the number of clusters per strata varied in proportion to the stratum population). Existing dwelling units in each selected cluster were listed, and each dwelling unit was scheduled to receive a screening interview. About 6600 families were contacted between December 1973 and late February 1974, and screening data were collected for 5800 families.

Baseline Interview Sample

Of those families who completed the screening interview, 4631 (79 percent) met the eligibility criteria for enrollment in the HIS:

1. At least one family head was between the ages of 18 and 59 (five-year sample) or 61 (three-year sample).
2. The weighted average of family income (0.4 times 1972 income plus 0.6 times 1973 income, all in 1973 dollars) was $25,000 or less.
3. No family head was on active duty or retired from any of the federal uniformed forces (mostly military) and eligible for free medical care or had a service-connected disability and was thus eligible for care through the Veterans Administration.

Using proportional stratification, about 2000 families were selected to receive the Baseline Interview. These interviews were completed during the summer and fall 1974 for about 1600 families (80 percent completion rate) representing some 4400 people.

Enrollment Sample

The Baseline Interview provided all the information necessary to select a sample to be offered enrollment. Families were assigned for possible enrollment in one of the HIS experimental health insurance plans, including a control group, using a complex statistical model, the Finite Selection Model (Morris, 1979). This model helped ensure that, across plans, families closely resembled each other on a variety of characteristics, including family income, age and race of family head, existence of previous health insurance, family size and number of family heads, education of family members, proportion of women in the family, medical care expenditures in the previous year, and health status.[1]

[1] Health status was measured in the Baseline Interview in terms of the respondent's rating of health as excellent, good, fair, or poor, and in terms of the amount of worry and pain caused by health. The measures described in other volumes of this series were not available or not used. One battery of

At the request of the Department of Health, Education, and Welfare, which funds the HIS, low-income families ($9000 or less annually) were slightly oversampled in Dayton (they had a 20 percent greater chance of inclusion), with a concomitant undersampling of middle-income families (between $9001 and $14,999 annually). Between October 1974 and February 1975, 646 Dayton families were offered enrollment in an experimental plan or in the control group, and 593 accepted (representing 1772 individuals); the acceptance rate was thus about 92 percent.[2]

DATA-GATHERING METHODS

The HIS has several major sources of data on the health status of individuals, including

Baseline Interview: an interviewer-administered questionnaire to determine whether a family was eligible for enrollment in the HIS. The Baseline Interview was completed in the respondent's home several months before possible enrollment. Respondents were heads of households; one head of the family could answer for the other if the latter was unavailable. Each family head received $5 for completing the questionnaire. The interview comprised several modules, including batteries of questions on health status, patient satisfaction with medical care, health insurance coverage, and medical care expenditures (hospital, physician, and dental). Information on sociodemographic and economic variables was obtained to update the eligibility information available from the screening interview.

The Dayton Baseline Interview included a maximum of 442 questions; skip patterns made the average number of questions answered difficult to calculate. Some questions pertained to the family as a whole, others to each family member. One-tenth of these questions pertained to health status (i.e., the first version of the Functional Limitations battery, a physical health measure). The average interview was completed in about two hours.

Enrollment Medical History Questionnaire: a self-administered questionnaire specific to three different age groups (14 and older, 5 to 13, and 0 to 4) completed in the respondent's home at enrollment. Adults completed the MHQ for children under 14. In Dayton, the enrollment MHQ had two parts: Form A was completed by (or for) all individuals at enrollment, and Form B was completed by a random subsample of enrollees selected to receive a medical screening examination (see below). In all other sites, both forms were completed by (or for) all enrollees at the time of enrollment. In Dayton, adults who completed Form A received a $2 compensation. Heads of families who completed both Forms A and B and the medical screening examination each received $20; for adult and child dependents, they received $5 per dependent for completing both forms and the examination, up to a family maximum of $50.

The MHQ was divided into two parts because of its length. Form A (ages 14 and older), which had a maximum of 531 questions, contained several batteries: the second version of the Functional Limitations battery, the HIS General Well-Being

physical health measures was included in this interview, but these scores were not used in the selection process.

[2] Approximately 630 individuals in 200 families participated in the control group during the experiment's first year in Dayton.

Schedule (a mental health measure), health habits, acute symptoms and conditions, several "tracer" conditions, life events and stress, efficacy of medical care, and other health-related information. Form B (ages 14 and older), which had a maximum of 388 questions, included the Physical Abilities battery, additional mental health items, and other "tracer" batteries.

Health Reports: a biweekly questionnaire, completed by the head of household (generally the female head), which covered the occurrence of restricted activity and bed disability days and use of medical and dental services on a person-by-person, day-by-day basis. The family was paid $4 for completing each report, which included approximately 12 health status questions for each family member.

Health Questionnaire: a self-administered questionnaire completed by the individual enrollee (or a parent as proxy respondent) annually, close to the anniversary date of enrollment. Batteries of items included in this questionnaire were identical with those in the MHQ; a maximum of 225 questions appeared on the adult (ages 14 and older) form. Families received $5 per family head for completing the questionnaires.

Exit Medical History Examination: a self-administered questionnaire similar in content to the enrollment MHQ. Adults completed the questionnaire for children under 14. All persons completed both Forms A and B and were compensated in conjunction with completion of the exit screening examination.

Medical Screening Examination: a series of physical examination and clinical laboratory procedures (e.g., audiometry, urinalysis) administered to adults and children in a randomly selected sample of families on enrollment (51 percent or 991 of 1772 in Dayton) and to all families at exit. The entire battery included some 20 different procedures and tests; the number administered to each enrollee depended on age, sex, and reported symptoms. Adults spent approximately two hours at the screening center, the time required to complete their examination and Form B of the MHQ for themselves and their children. The average time required for pediatric and infant examinations was 45 minutes. (For further detail, including descriptions of the screening examination procedures, see Smith et al., 1978.)

Structured response choices, rather than open-ended questions, were used for all health status items in the preceding questionnaires. Questionnaires were checked for missing items. In administering the Baseline Interview, interviewers encouraged applicants to answer every question. For the self-administered MHQ, if answers were lacking for more than six items, respondents were contacted by telephone to obtain the missing information. If the respondent had problems with vision or understanding the questions, the usually self-administered questionnaires were interviewer-administered, and the difference in administration noted in the data base. Data were processed using standardized coding procedures and then "cleaned" by a computer program that checks for possible coding errors and assigns a data status indicator describing the quality of data for each item in the questionnaire.

Most of the information presented in the R-1987-HEW series was based on analyses of data from the Dayton enrollment MHQ; some data reported in this series came from the Baseline Interview. The conditions under which these instruments were administered are certain to have influenced conclusions regarding data quality (response rates, reliability, validity, and power) as well as estimates of central tendency and variability. Therefore, these methodological details should be kept in mind when interpreting HIS results or attempting to generalize them to other populations or research settings. To facilitate this process, a summary list of factors that may influence survey results is presented in Table 4, along with a brief description of how each was handled in Dayton when the Baseline Interview and MHQ were administered.

RESPONDENT CHARACTERISTICS

Analyses reported in the R-1987-HEW series were based on data collected from adult enrollees before the experiment began in Dayton. The Baseline Interview provided data for the first version of the Functional Limitations battery, and Forms A and B of the enrollment MHQ provided data for other physical health and all mental health batteries. Because batteries containing items explicitly constructed to measure social health and general health perceptions were not included on the Dayton enrollment MHQ, the R-1987-HEW series discusses only their conceptualization in the HIS, and presents no results based on Dayton data. Characteristics of respondents who completed these questionnaires are noted below.

Baseline Interview

Information on 3444 persons ages 14 and older was obtained on the Baseline Interview (see Table 5, first two columns). The average age of the sample was 35.3 (standard deviation, 15.4); the range was 14 to 90. Forty-eight percent were men and 52 percent women. Approximately 12 percent of the sample was nonwhite. The average number of school years completed was 12.3 (standard deviation, 2.8). Reported annual family income in 1973 ranged from $0 to $56,000, with a mean of $13,871 (standard deviation, $7799).[3]

Medical History Questionnaire, Form A

Most of the results presented in the R-1987-HEW series were based on data obtained in Dayton from 1209 enrollees ages 14 and older who completed Form A of the MHQ. (The number of adults completing this form was 1212; however, at the time most of the analyses reported in these volumes were performed, data on one family of four, one of whom was under 14, were missing.) The second pair of columns in Table 5 summarizes the characteristics of this sample. The average age was 34.3 (standard deviation, 14.2); the range was 14 to 75 (only 4.1 percent were over 60). Forty-seven percent were male and 52 percent female. Approximately 11 percent were nonwhite. The average number of school years completed was 12.6 (standard deviation, 2.9). Reported family income in 1973 ranged from $0 to $27,000, with a mean of $13,687 (standard deviation, $6375).

Medical History Questionnaire, Form B

Other analyses (specifically, those for the Physical Abilities battery reported in Vol. II and several mental health scales reported in Vol. III) were based on an

[3] The income variable used for the analyses reported here was the 1973 annual family income, not the weighted 1972-1973 average used to determine eligibility.

Table 4

Summary of Methodologic Factors That Influence Survey Results and Characteristics of Health Insurance Study Data-Gathering Methods

		Enrollment Medical History Questionnaire	
Factors	Baseline Interview	Physical Health	Mental Health
Context or purpose of data collection	To obtain information for determining eligibility for enrollment	Required of respondents enrolled in HIS	Required of respondents enrolled in HIS
Method of administration	Face-to-face interview	Self-administered	Self-administered
Interviewer characteristics	Highly trained interviewer	Self-administered	Self-administered
Location of administration	Respondent's home	Respondent's home (Form A), mobile clinic (Form B)	Respondent's home (Form A), mobile clinic (Form B)
Respondent interest in subject	Probably high, given focus on health status and health care delivery	Probably high, given focus on health status and health care delivery	Probably high, given focus on health status and health care delivery
Sensitivity of questions	Physical health in terms of functioning, probably one of the least sensitive aspects of health status	Physical health in terms of functioning, probably one of the least sensitive aspects of health status	Mental health, probably one of the more sensitive aspects of health status
Position of questions and length of questionnaire	Functional Limitations: 8 items administered following approximately 58 items[a]	Functional Limitations: Nos. 9 to 36 of 294 items; Physical Abilities: Nos. 157 to 166 of 199 items	HIS General Well-Being: Nos. 230 to 251 of 294 items; Form B battery: Nos. 167 to 199 of 199 items
Form of questions	Structured response choices	Structured response choices	Structured response choices
Compensation	$5 per head of family	$2 per adult (Form A)[b]	$2 per adult (Form A)[b]
Procedures to aid recall	None	None	None
Recall period	Functional Limitations: 3 months	Functional Limitations: 3 months; Physical Abilities: present health status	Past month
Instrument complexity	Moderate; forced choice responses to items and skip patterns involved	Moderate; forced choice responses to items and skip patterns involved	Low; forced choice responses to items and no skips
Abstractness of concept	Physical health in terms of functioning is less complex, less abstract than other aspects of health	Physical health in terms of functioning is less complex, less abstract than other aspects of health	Mental health constructs tend to be more complex and more abstract than other aspects of health
Population group	Sample of Dayton population: age 14 to 90 (mean=35); 48% male; 85% white; mean educational level 12.3 years; mean income $13,865 (1973 dollars)	Sample of Dayton population: age 14 to 75 (mean=34.3); 48% male; 87% white; mean educational level 12.6 years; mean income $13,687 (1973 dollars); low income ($9000 and under) and younger than 61 oversampled	Sample of Dayton population: age 14 to 75 (mean=34.3); 48% male; 87% white; mean educational level 12.6 years; mean income $13,687 (1973 dollars); low income ($9000 and under) and younger than 61 oversampled
Use of proxy respondents	Rare	Rare	Rare
Field edit specifications	Careful editing for missing items; call back when more than 6 missing	Careful editing for missing items; call back when more than 6 missing	Careful editing for missing items; call back when more than 6 missing
Data preparation methods	Standardized coding procedures	Standardized coding procedures	Standardized coding procedures

[a] Skip patterns and family size determined the actual number of items completed before administration of the Functional Limitations battery; the average respondent filled out 58 items.

[b] Each adult (over 14) who completed Form A received $2 and each child $0.50; in families who completed both Forms A and B (most of whom also received the multiphasic screening examination), each head received $20 and each dependent $5, up to a maximum of $50 per family unit. These figures differed at enrollment in sites other than Dayton.

Table 5

Characteristics of Adult Sample (Ages 14 and Older) Completing Baseline Interview and Enrollment Medical History Questionnaire, Dayton

	Questionnaire					
	Baseline Interview (N=3444)		MHQ, Form A (N=1209)		MHQ, Form B (N=835)	
Characteristic	N	Percent	N	Percent	N	Percent
Age group						
14-19	635	18.4	221	18.3	160	19.2
20-24	412	12.0	134	11.1	96	11.5
25-29	448	13.0	169	14.0	104	12.5
30-34	356	10.3	137	11.3	82	9.8
35-39	278	8.1	111	9.2	75	9.0
40-44	284	8.2	99	8.2	76	9.1
45-49	282	8.2	89	7.4	62	7.4
50-54	283	8.2	114	9.4	79	9.5
55-59	220	6.4	72	6.0	55	6.6
60 and older	246	7.1	50	4.1	37	4.4
Missing	0	0.0	13	1.0	9	1.1
Race[a]						
White	2327	84.5	852	88.4	579	88.0
Black	398	14.4	105	10.9	72	10.9
Other	20	0.7	7	0.7	7	1.1
Missing	8	0.3	0	0.0	0	0.0
Sex						
Male	1655	48.1	569	47.1	392	46.9
Female	1789	51.9	627	51.9	434	52.0
Missing	0	0.0	13	1.1	9	1.1
Education[b] (years)						
< 8	106	3.5	31	3.0	24	3.4
8-11	660	22.1	209	20.4	147	20.8
12	1203	40.3	396	38.7	277	39.2
13-16	778	26.0	294	28.7	203	28.8
17 and over	178	6.0	85	8.3	48	6.8
Missing	62	2.1	9	0.9	7	1.0
Income[c]						
< $9000	896	26.0	303	25.1	211	25.3
$9000-$14,999	1027	29.8	336	27.8	218	26.1
$15,000 and over	1463	42.5	557	46.1	397	47.5
Missing	58	1.7	13	1.1	9	1.1

[a] Obtained for adult heads of household only; thus, relevant samples are smaller.

[b] Obtained for respondents ages 18 and older only; thus, relevant samples are smaller.

[c] 1973 family income; eligibility criteria restricted range to under $25,000 for MHQ samples.

income-stratified random sample of 835 adult enrollees who completed Form B of the MHQ (in addition to Form A). The right-hand columns of Table 5 summarize the characteristics of this subsample. The average age was 34.6 (standard deviation, 14.5); the range was 14 to 75 (only 3.2 percent were over 60). Approximately 47 percent were male and 52 female. Some 11 percent were nonwhite. The average number of school years completed was 12.5 (standard deviation, 2.8). Reported family income in 1973 ranged from $0 to $27,000, with a mean of approximately $13,848 (standard deviation, $6572).

CONSTRUCTION AND EVALUATION OF MULTI-ITEM SCALES

Several methods were used to construct multi-item scales from adult questionnaire items fielded in Dayton and to evaluate them as measures of health status. This section describes the methods used (1) to combine questionnaire items pertaining to the same health status construct into multi-item measures; (2) to assess variability of scores; (3) to determine how much information rather than random error was obtained (to study reliability of measurement); (4) to ascertain the extent to which each measure assessed the particular health status construct it was intended to measure (to study validity of measurement); and (5) to determine the extent to which it will be possible to reject the null hypothesis that no differences in health status are produced by different HIS insurance plans if such differences do exist (to study power of measurement). (The philosophy and formulas of power analysis are presented only briefly in this volume. They are discussed at length in Vol. VII, which is devoted entirely to results of power analyses of the physical and mental health measures fielded in Dayton; power is not discussed in other volumes in this series.)

An important adjunct to studying the adequacy of HIS measures were reviews of the literature on measurement of each health dimension selected for the HIS—physical, mental, and social health and general health perceptions. These reviews were performed to identify issues that needed to be addressed when selecting from among measures published by others and when constructing new HIS measures. They focused on developments in the state of the art of measuring each of these dimensions over the past 30 to 40 years. Articles were identified by reviewing files maintained by HIS staff, which contain 500 to 600 publications that are updated by periodic screening of some 25 journals that frequently publish articles on conceptualization and measurement of health status variables.[4] The literature reviews also provided a framework for better understanding the strengths and shortcomings of HIS health status measures.

Item Scoring and Scaling Methods

Questionnaire items can be scored in several different ways to define variables for analytic use. Relying on a single item to assess each variable is least desirable,

[4] These journals include *American Journal of Psychiatry, American Sociological Review, British Journal of Preventive and Social Medicine, Educational and Psychological Measurement, Health Services Research, Health and Society/Milbank Memorial Fund Quarterly, Inquiry, International Journal of Epidemiology, Journal of Abnormal Psychology, Journal of Clinical Psychology, Journal of Community Mental Health, Journal of Health and Social Behavior, Medical Care, Social Indicators Research,* and *Social Science and Medicine.*

because of relatively poor reliability, validity, and power of measurement. Using several single items scored separately for each variable enhances these measurement properties, but univariate analyses of highly correlated items measuring the same variable are inappropriate. Constructing multi-item measures to assess each analytic variable has several advantages when compared with the use of single-item measures.

For HIS analyses, multi-item measures were used because they (1) reduce the number of scores necessary to define each health status variable; (2) increase score reliability by pooling the information that items have in common; (3) increase validity (if items are carefully selected to provide a more representative sample of information about the construct); (4) minimize bias caused by tendencies to endorse or negate items regardless of content (in cases where both favorably and unfavorably worded items are combined); and (5) provide the option, if item responses are missing, to estimate responses using other items in the measure, thus reducing missing scores on the multi-item scale.

Several assumptions underlie the combination of items into multi-item measures of health status constructs: (1) each item contributes to the pool of information about the construct; (2) the method of combination is consistent with the relationships among the items; (3) all items are scored in the same direction (e.g., high scores indicate good health); (4) all items measure the construct in approximately the same units (or the items have been standardized and weighted); and (5) the resulting multi-item measure has a meaningful interpretation.

Empirical scaling analyses carried out during studies of HIS health status measures were designed to make sure these assumptions were fulfilled. The first step in these analyses was to hypothesize groups of items that could be combined into a single score. Hypotheses were based on results reported by other investigators who had used the same or similar items, or on logical combinations of items appearing from their content to measure the same construct. Two different models of relationships among items and their related scaling methods were used to test the appropriateness of item groupings: Likert's (1932) Method of Summated Ratings and Guttman's (1944) Scalogram Analysis. The Method of Equal-Appearing Intervals (Thurstone and Chave, 1929) was considered and rejected; scales developed according to this method are similar in reliability and validity to those constructed using the other methods but would have been more costly because a panel of judges is required to derive scale values for items before constructing and fielding scales (see Edwards, 1957; Fishbein, 1967).

Method of Summated Ratings. A modified version of Likert's Method of Summated Ratings was used to score one battery of physical health measures (Physical Abilities), the mental health scales, and the general health perceptions scales (see Vols. II, III, and V, respectively).[5] This method is based on two assumptions regarding relationships among items: (1) that items in each hypothesized grouping all contain the same proportion of information about the same health construct(s) (i.e., items need not be weighted for differences in factor content when they are combined); and (2) that items contribute equally to the variance of the total scale score (i.e., that they have approximately equal variances and thus item scores need not be standardized). This scaling method works best when items from the

[5] Scaling analyses of the general health perceptions battery fielded after Dayton were based on data from non-HIS general populations.

same scale sample all important aspects of the same construct and when the response scale for each item provides for a range of responses. For example, one item in the mental health battery assessed the frequency of feeling "depressed" from all to none of the time, and another the frequency of "feeling downhearted and blue" from all to none of the time; in this example, the items assess two highly related but distinct aspects of mood or affect and each provides for a full range of responses on the mood/affect continuum.

Construction of HIS summated ratings scales from item batteries followed five steps. These steps added several scaling criteria to those usually associated with Likert scaling. Performed during HIS scaling studies in the order shown below, they were designed to determine whether

1. Each item in a hypothesized grouping was substantially linearly related to the total score computed from items in that group (traditional Likert criterion).
2. Each item correlated higher with the construct it was hypothesized to measure than with other constructs (item discriminant validity criterion).
3. Item groups not hypothesized a priori might be identified (factor analytic test).
4. Items in the same scale contained the same proportion of information about the same constructs or should be given different weights (test for equal factor loadings).
5. The score for each item required standardization before combining it with other items in the same scale (equal variances criterion).

If items in each hypothesized grouping roughly satisfied these assumptions, responses to the items were simply summed to derive a scale score for that construct. If numerous scaling errors were encountered in a priori hypothesized groupings, the hypotheses were restated on the basis of empirical findings, and revised item groupings were reevaluated according to these criteria. In those few instances (e.g., Form B mental health items described in Vol. III) when item groupings were not hypothesized in advance, they were based on manifest item content, and the structure of the battery was explored using factor analysis.

Multitrait Scaling. The first two steps used to construct summated ratings scales were performed during multitrait scaling analyses and involved examining a matrix of item-scale correlations.[6] Constructs were operationally defined as scale scores, and a matrix of product-moment correlations was computed. Each row of the matrix contained correlations between scores for one item and all hypothesized item groupings (constructs defined by scales). Each column contained correlations between the scores for one scale and all items, including those hypothesized to be part of that scale and those hypothesized to be part of other scales. Correlations between each scale and items used to score that scale were corrected for overlap[7] so that estimates of the item-construct relationships were not spuriously inflated.

[6] All computations were performed using the ANLITH (Analysis of Item-Trait Homogeneity) program, which was written by Thomas Gronek at IBM and Thomas Tyler at the Academic Computing Center at Southern Illinois University, and was modified for use at Rand in HIS scaling analyses.

[7] The method used for correcting item-scale correlations for the effect of relevant item inclusion was that suggested by Howard and Forehand (1962). An item-scale correlation corrected for overlap is the correlation between the item score and the sum of items in that scale other than the item in question. Such a correction represented a modification of the traditional Likert criterion and was made because the number of items in each group was much smaller than has been common for Likert scales, and thus each item had a considerable influence on the total scale score.

The first step in the analysis of these matrices involved examining the magnitude of item-scale correlations. The multitrait matrix was inspected to verify that each item-scale correlation was substantial (i.e., about 0.30 or higher, absolute magnitude). Any item not having a strong linear relationship (corrected for overlap) with the total score for the scale that included the item did not meet the Likert-type criterion and was eliminated from that scale. Some items eliminated at this stage were retained for separate analysis pending tests of whether they should be placed in another scale.

For the second step, the highest correlation in a row should be the one between the item and the scale defining the construct it was hypothesized to measure. This step was a test of discriminant validity (following Campbell and Fiske, 1959). The item discriminant validity criterion was satisfied and a scaling "success" counted whenever the correlation between an item and its hypothesized scale was more than two standard errors higher than other correlations in the same row. When a correlation between an item and its hypothesized scale was more than two standard errors below another correlation in the same row, a "definite" scaling error was counted. When the correlation between the item and other scales in the same row was within two standard errors of its correlation with its hypothesized scale, a "probable" scaling error was counted. In cases of this last type, it was not possible to predict whether correlations between the item and its hypothesized scale would be higher or lower than others in the same row upon replication of the analysis. To take these marginal results into account, a "probable" scaling error was counted.

The number of discriminant validity tests for each scale equaled the number of items in that scale times one less than the number of scales in the matrix. "Definite" errors were rare. "Probable" errors were concentrated in the same items, making it fairly easy to identify appropriate scale revisions (e.g., eliminating problematic items).

Table 6 presents a sample item-scale correlation matrix, drawn from a larger matrix evaluated during analyses of HIS mental health measures (see Vol. III). The first scaling criterion, a traditional Likert-type criterion, requires a substantial item-scale correlation (corrected for overlap). Correlations between items and hypothesized scales (i.e., those identified by asterisks in Table 6) must exceed 0.30 to satisfy this criterion. In this example, they do so without exception. The second criterion, item discriminant validity, requires that the correlation between each item and its hypothesized scale be two standard errors higher (about 0.06 higher in this example) than its correlation with other scales. For example, the first Anxiety item correlated 0.67 with its hypothesized scale and from 0.52 to 0.56, absolute magnitude, with other scales. Thus, two scaling "successes" would be counted for that item. No "definite" scaling errors were observed for these items. One "probable" error each would be counted for the first and last Depression items; correlations between each of these items and other scales were within two standard errors of its correlation with the hypothesized scale.

Factor Analysis. For the next two steps in constructing summated ratings scales, factor analysis was used to test for unhypothesized item groupings and to evaluate factor loadings for items.[8] In a factor analysis, the factors identified repre-

[8] See Cronbach (1970), Comrey (1967, 1973), or Armor (1974) for further information regarding factor analysis and homogeneity of measurement.

Table 6

SAMPLE ITEM-SCALE CORRELATION MATRIX: CORRELATIONS AMONG SELECTED MENTAL HEALTH ITEMS AND HYPOTHESIZED SCALES

Item Grouping/Abbreviated Item Content	Hypothesized Scales		
	ANX	DEP	PWB
Anxiety (ANX)			
Bothered by nerves	67*	56	-52
Strain, stress	71*	56	-50
Anxious, worried	73*	70	-55
High-strung vs. relaxed	72*	58	-60
Tense, tension	78*	60	-57
Depression (DEP)			
Depressed	64	69*	-59
Sad, discouraged	59	72*	-57
Downhearted, blue	67	72*	-63
Positive Well-Being (PWB)			
Feeling in general	-64	-62	69*
Happy with life	-54	-58	67*
Daily life interesting	-40	-46	59*
Cheerful, lighthearted	-59	-57	68*

NOTE: Based on data reported in Table 14, Vol. III; decimals have been omitted.

*Item-scale correlation is corrected for overlap (i.e., correlation between item and sum of all other items in the scale).

sent underlying dimensions of measurement defined by the items. The multitrait scaling tests were based on a particular hypothesized structure underlying the battery (i.e., constructs were necessarily defined by groups of items in advance). Factor analysis was used to test for unhypothesized factors that could account for scaling errors in the multitrait analyses. Factor analyses were also performed to test whether weights (factor loadings) were comparable across standardized items in the same scale. If so, each item could be given the same unit weight.

In nearly all cases, the factors corresponded to constructs defined in the a priori hypothesized item groupings. When this occurred, items defining each factor were examined to determine whether they were identical with those used to define hypothesized scales in the multitrait scaling tests. Only items that correlated substantially with the same factor were retained in that item grouping. In some cases, factor analysis identified item groupings that were not identical with those hypothesized before multitrait scaling studies. When this happened, the differences tended to correspond to the pattern of multitrait scaling errors and provided additional insights on how to better define the scales.[9]

Principal components analysis[10] was chosen as the method of factor analysis. For each battery of items, factors were extracted from a matrix of product-moment correlations among item scores with unities in the matrix diagonal; that is, all of the variance in the items was to be explained by the factors. Use of reliability estimates in the diagonal would have been preferable, but they were not available when the factor analytic studies were carried out. Insertion of communality estimates (the amount of variance an item has in common with other items in the matrix) in the diagonal was considered unsatisfactory, because factors represented by only one item would be more likely to be lost.[11]

The first factor identified in principal components analysis is the most reliable one defined by the items.[12] If only one factor is identified using the criterion that factors must be associated with eigenvalues (sum of squared factor loadings) equal to or greater than unity and if all items measure that factor equally, homogeneity of measurement and simple scoring methods are supported for the items. If more than one factor is extracted or if correlations between items and a given factor are unequal, more than one scale score may be required to adequately represent the constructs defined by the items, or some items should not be used to compute scale scores.

Before rotation, the initial (unrotated) solution was evaluated against several criteria. In their study of different factor analytic methods applied to the same batteries of items, Ware, Miller, and Snyder (1973) determined that application of these criteria resulted in the same or very similar scaling decisions regardless of the method of factor extraction (e.g., principal factor, principal components, image analysis). The criteria used in HIS analyses were the following:

- The Scree Test (Cattell, 1966), which involves interpreting the curve relating the factors to the proportion of total variance accounted for by each factor. The test is based on the assumption that when the decreasing negative slope of the curve begins to level off, random error factors have been encountered.[13]
- The 5 percent guideline described by Guertin and Bailey (1970), which suggests that factors associated with 5 percent or more of the common variance warrant further study.
- Identification of true common factors, in which only unrotated factors having two or more loadings of 0.30 or greater (absolute magnitude) are selected for rotation and interpretation.
- Use of trial rotations when the decision as to the "best" number of factors for final rotation is ambiguous according to the preceding criteria. Trial

[9] When factor analysis is used in this way, empirical findings on item groupings that are not hypothesized in advance must be replicated or cross-validated in an independent sample before hypotheses and operational definitions can be modified with confidence. The scaling studies of revised batteries using data from other HIS sites will prove useful for this purpose.

[10] The SPSS program PA1 was used (Nie et al., 1975).

[11] See Guertin and Bailey (1970).

[12] See Cronbach (1951) or Armor (1974) for a discussion of the relationship between factors and internal-consistency reliability.

[13] As noted by Ware, Miller, and Snyder (1973), interpreting the Scree curve involves a certain amount of subjectivity.

> rotations are evaluated in terms of interpretability and the meaningfulness and desirability of alterations in major factors when additional factors are rotated.

In the final rotation, each factor should become more clear when rotated to an orthogonal simple structure without disrupting other major factors.

Once final item groupings were identified, the last step involved comparing item variances to determine whether standardization was required before scoring the scale. If correlations between all items in a group and the principal component identified in the factor analysis were comparable, each standardized item was given the same weight in scoring the scale. Further, if standard deviations were equivalent, item scores did not need to be standardized before scoring the scale. Thus, each summated ratings scale could be scored by computing the simple algebraic sum of response scores for those items that satisfied multitrait and factor analytic scaling criteria.

Missing Data. When summated ratings scales were scored, substitutions had to be made for missing item responses—typically less than 1 percent of possible responses. Several options were considered, including

1. Midpoint of the possible scale range.
2. Sample central tendency statistics: mean, median, or modal score for the item in question.
3. Respondent central tendency statistic: mean, median, or modal score for that respondent across either all items in the battery or other items in the same scale. When the range of response values differs for items used (e.g., one item with four possible responses and another with five), responses were prorated to estimate the missing response.
4. Regression estimate.

These options ranged from one that could be implemented with few resources and could be the least accurate (option 1) to the most costly and most accurate (option 4). Considering the tradeoffs involved, which are discussed below, the third option was selected to handle missing data when constructing HIS summated ratings scales.

The first option substitutes the middle of the possible score range for the missing response, and makes it possible to program steps for estimating such responses before any statistical analysis. The method can be effective when items have symmetrical distributions and central tendencies close to the middle of the possible score range and a homogeneous sample is being studied. When score distributions are skewed or respondents vary greatly, estimated scores can differ greatly from actual scores with this method.

When the midpoint of the possible scale range and central tendency statistics do not agree, the second option—substitution of a central tendency estimate based on data from the entire sample—will usually provide better estimates. This option does not focus, however, on information about the respondent in question; such information is usually available either from other items in the battery or from those in the same scale. If available responses are to be used in estimating missing data, information from the respondent—the third option—should prove more useful than information gathered from all respondents. One disadvantage of this option is the cost of computing, for example, respondent mean scores that would otherwise not be computed. In addition, use of respondent means ignores differences in relationships between the missing items and items used to estimate missing scores. This produces greater estimation error, on average, than if an optimal weight were derived for each available item.

Use of regression estimates for missing responses accounts for the possibility that available scores for other items may be weighted unequally. Option 3 assumes that each item correlates perfectly, or at least at the same magnitude, with the scale total. In practice, this assumption is rarely met. When departures from this assumption are substantial, the regression method of estimating missing scores represents an improvement. Using the regression method, optimal weights can be derived by regressing each item on other items in the same scale using data for the total sample. Scores estimated using this method cannot deviate from true scores more, on average, than those estimated using the three methods discussed above. Although it yields the most accurate estimates, this option is also the most costly; resources used to estimate scores deplete those available to test hypotheses. Choice of the regression option should be made on the basis of whether gains in accuracy warrant the increased cost and complexity of scoring missing data.

In the HIS, the first two options were rejected. Item response distributions for batteries used to score summated ratings scales were often skewed, decreasing the appropriateness of the first option. Further, considerable variability in individual responses was observed, suggesting that responses to other items by the same respondent would be more useful than the responses of others to the item in question.

The methods used to construct HIS summated ratings scales group together items having similar relationships to the construct being measured. Thus, these items tend to have similar variances and to contribute equivalently to the score on the measured construct. Use of a central tendency estimate based on the respondent's answers to other items in the scale (option 3) should therefore yield estimates similar to those provided by the regression method (option 4). For single-factor scales, missing item scores were estimated by computing a prorated scale score from available item responses in that scale; the mean score for available items was the estimate of the score for the missing item(s). When responses to all items in the same scale were missing for a respondent, the scale score was coded as missing.

Guttman Scalogram Analysis. Scalogram Analysis (Guttman, 1944), or Guttman scaling,[14] was used to construct multi-item measures from items in the functional status battery. This method of scale construction assumes that each item is designed to assess a specific level of the construct being measured and that the pattern of responses across items is reproducible from the scale score. In addition to evaluating the extent to which items in the same grouping measure the same construct (are unidimensional), Scalogram Analysis evaluates whether items are correctly ordered by level (e.g., level of severity of limitation) and whether the hypothesized pattern of scores across items can be reproduced from the scale score (whether a cumulative scale is defined). HIS items subjected to Scalogram Analysis were each designed to measure a different point along an ordered continuum. Thus, several items were necessary to define the entire range of the health status construct being measured.

[14] See Edwards (1957) or Jackson and Messick (1967) for further discussions of Guttman scaling methods.

For example, three items may be used to assess a person's physical health status in terms of physical capacity:

1. Able to run a mile.
2. Able to walk a mile.
3. Able to walk a block.

If each item is unidimensional and assesses a different level of ability to get around and if the scale is cumulative, each respondent should answer these items according to one of several logical patterns of responses:

	Item and Response		
Scale Score	Run a Mile	Walk a Mile	Walk a Block
3	Yes	Yes	Yes
2	No	Yes	Yes
1	No	No	Yes
0	No	No	No

Logically, a person able to run a mile should also be able to walk a mile and a block; someone unable to run or walk a mile might or might not be able to walk a block. Other response patterns, such as being able to walk a mile but not walk a block, would not be logical.

To scale a group of items using Guttman scaling, each item must be scored dichotomously; in the HIS, responses of yes and no were used. Scores of 1 and 0 were assigned, respectively, to these responses if the item represented a favorable definition of health (as do all items in this example); scores of 0 and 1 were assigned if the item represented an unfavorable definition (e.g., confined to bed all or part of the day).

In addition, two or more items hypothesized to measure different levels may be combined if they appear empirically to define the same scale type. If this is done, the combined item must also be scored dichotomously. For example, a limitation in functioning may be scored at the same level if a person is unable to perform certain kinds of housework or schoolwork.

The number of items endorsed represents the scale level that defines the respondent's health status. In the preceding example, a scale score of three indicates that the respondent was able to perform all of the activities described, while a scale score of zero indicates that the respondent was unable to perform any of these activities. Two individuals will have the same score on a perfect Guttman scale only if their responses are identical on all items. (This is not the case with summated ratings scales, on which the same score may be obtained by different patterns of answers to items in the scale.)

Not all respondents, however, answer items in a manner consistent with one of the hypothesized patterns. Any deviation from a hypothesized pattern is counted as a scaling error. If the number of errors observed in the analysis is small in relation to the total possible number of errors, a reproducible scale has been achieved.

Two coefficients were used to evaluate whether items met Guttman criteria (i.e., their scalability). First, the coefficient of reproducibility, CR = 1 − (observed errors/total possible errors), was computed. All deviations from hypothesized response patterns were initially counted as observed errors; the total possible number of errors was N (the number of observations) times the number of items being scaled. A high CR value indicates both reliability (in the internal-consistency sense) and reproducibility (see further discussion under "Reliability" below). Following guidelines suggested by Guttman (1944) and Edwards (1957), CR values of 0.90 or greater were accepted as evidence of the reliability and reproducibility of a battery of items.

It was also necessary to evaluate the extent to which each observed CR represented an improvement over its minimum possible value, because CR can be large even when a truly cumulative scale is not achieved. For example, if 90 percent of the respondents indicate no limitations across all items (a scale score of zero), CR would equal or exceed 0.90 regardless of the pattern of responses observed for persons with other scale scores. For this reason, the extent to which the observed CR represented an improvement over the minimum possible CR was examined. This minimum marginal reproducibility (MMR)[15] is the smallest possible value of CR given the distributions of item responses. CR and MMR were compared using the coefficient of scalability, CS = (CR − MMR)/(1 − MMR), which indicates the proportion of possible improvement in MMR that was achieved by the scale. The recommended standard for CS of 0.60 was accepted as evidence of scalability (Nie et al., 1975). Because CR values are artificially inflated if item score distributions are highly skewed, HIS Guttman Scalogram analyses emphasized CS when evaluating the scalability of items and were performed after eliminating the large number of HIS enrollees who received perfect scores across all items.

Missing Data. To estimate missing responses for items in Guttman scales, items were reviewed on a case-by-case basis. (The number of missing items was small for HIS analyses, typically a fraction of 1 percent.) Inspection of the total pattern of responses across completed items in a given scale allowed a "best guess" regarding the appropriate scale level to be assigned the respondent; substitute values were not assigned to individual items. For example, if someone reports being able to run short distances and does not respond to an item about walking a block, it would be assumed the respondent could walk a block.

Descriptive Statistics for Scales (Studies of Variability)

After empirical scaling studies were completed for HIS health status measures, the variability of resulting score distributions was studied. Score distributions should adequately represent the actual distribution of health status on the particular dimension or construct being measured. If this is achieved, the measure can be

[15] CR has no unique minimum value and is a function of the marginal frequency for each item in the scale. When the distributions of these frequencies depart from 50 percent for dichotomous items, the lower limit of CR rises sharply (from its absolute lower limit under perfect conditions). The chance probability of each scale type is the product of the proportions of persons making correct responses to each item. With skewed item distributions and short scales such as those constructed in the HIS, the chance probability of each scale type and the MMR can be very high (even exceeding the standard of 0.90) or very low (almost zero). See White and Saltz (1967) and Schooler (1974) for further information.

used to detect differences in health status in the specific population whose health status is being assessed.

Variability of scores on health status measures may be insufficient for many reasons. Assuming there is actual variability in the construct being measured, insufficient variability may indicate that the items (1) do not adequately assess the particular health construct of interest; (2) do not adequately detect differences in some range of values between the extremes (e.g., distances between the levels represented by the items may be too large, and scores may not reflect important differences between the health states of respondents); and (3) do not assess important differences in health states at one or the other end of the continuum (e.g., items assessing severe functional limitations only). In any of these cases, potential effects of the experiment may be missed. Addition of items that assess clinically significant differences between scale levels more precisely and that increase the range of measurement should increase the variability of resulting score distributions and the usefulness of the scale in detecting actual differences in health status.

Studies of score variability on HIS health status scales focused on these three issues. The shape of distributions was also inspected to identify measures yielding skewed score distributions. Those that were fairly normal, or at least roughly symmetrical, will prove more powerful for purposes of statistical analysis.

Not all changes in questionnaire items designed to reduce coarseness or to extend the range of measurement will necessarily improve precision for purposes of testing hypotheses in the HIS. For example, if measurement were improved in areas that cannot be affected by increased access to medical care, such improvement would not change the ability to detect differences attributable to health care financing. This consideration was kept in mind when revising HIS items.

Studies of Reliability

Reliability of measurement refers to the extent to which measured variance reflects true score rather than random error. Reliability is a prerequisite to use of a score for any purpose. A reliability coefficient is an estimate of the proportion of total variance that is true score variance, as expressed in the following formula (from Kerlinger and Pedhazur, 1973):

$$\text{Reliability} = 1 - \frac{V_e}{V_t} ,$$

where V_e equals the error variance and V_t is the total measured variance. Reliability can be estimated in a variety of ways, which differ in underlying assumptions and methods. Three methods were used to study HIS health status measures: internal-consistency, test-retest, and reproducibility. An excellent discussion of the different assumptions underlying internal-consistency and test-retest methods is presented by Cronbach (1951). The relationship between reliability and reproducibility is discussed by Jackson and Messick (1967).

Internal-Consistency Reliability. The internal-consistency method of estimating reliability applies to multi-item scales. The reliability coefficient it yields is a function of two properties of scale items: (1) item homogeneity, or the extent to which the items share common variance; and (2) the number of items in the scale. The relationships among internal-consistency reliability, homogeneity, and scale length are shown in the following formula (from Nunnally, 1967):

$$r_{tt} = \frac{kr_{ii}}{1 + (k - 1)r_{ii}} ,$$

where r_{tt} is the internal-consistency reliability of a score, k is the number of items used to compute the scale score, and r_{ii} is the average inter-item correlation (Fiske, 1966; Tyler and Fiske, 1968). Reliability can be increased by lengthening a heterogeneous scale or by selecting some chosen number of items that are more homogeneous.

To compare the reliability of scales, both the internal-consistency reliability and homogeneity of measurement are of interest. In scales that contain the same number of items, r_{tt} indicates both reliability and homogeneity. For scales that differ substantially in length, reliability can be compared by using an estimate (r_{ii}) that is not affected by the number of items included in the scale. Both r_{tt} and r_{ii} were reported for HIS measures.

Because several volumes in the R-1987-HEW series discuss reliability findings reported in the literature and because many investigators reported only r_{tt}, r_{ii} was calculated from reported information using the following formula (from Cronbach, 1951):

$$r_{ii} = \frac{r_{tt}}{k - (k - 1)r_{tt}} ,$$

where r_{ii}, r_{tt}, and k are the quantities defined earlier.

The internal-consistency approach was used to estimate reliability for HIS scales constructed using the Method of Summated Ratings (e.g., scales to measure physical abilities and mental health). Internal-consistency estimates were considered acceptable for HIS purposes—to make group comparisons—if they were 0.50 or above. Coefficients of 0.90 or greater would be acceptable for individual comparisons (Helmstadter, 1964), which are not required for HIS analyses.

Test-Retest Reliability. The second approach used to estimate score reliability is applicable to both single- and multi-item measures. This approach requires repeated administrations of the same instrument and is based on the logic that people should receive the same score across observations using the same method if there is no change in the trait being measured. In the HIS, test-retest reliability was estimated by computing product-moment correlations between scores for the same respondents at two points in time. The test-retest method was used for one battery of physical health measures (functional limitations), which included one single-item measure, and for several mental health scales.

Scores for the functional limitations scales were computed from alternate forms of the battery included on the Baseline Interview and on Form A of the MHQ, which were administered approximately four months apart on average. Because this battery of items assessed chronic functional limitations (those present for more than three months), it is not likely that real changes in status between administrations attenuated estimates of reliability.

Scores for the mental health scales were computed from comparable batteries of items included on Forms A and B of the MHQ, administered within four weeks for most respondents and as long as four months apart for others. Mental health items defined a one-month recall period, and mental health constructs are probably less stable over time than chronic functional limitations. Thus, for many respondents, particularly those with a longer test-retest interval, coefficients for mental health measures were attenuated because they reflect both reliability of measurement at each administration and the stability of the mental health construct measured. For this reason, the length of time between administrations was controlled for in analyzing test-retest reliability for mental health scales. Differences between items across alternate forms of both the mental health and functional limitations batteries might also have attenuated test-retest coefficients, particularly if these differences meant that slightly different constructs were measured by the two forms.

Because HIS questionnaires are fielded in diverse populations that differ widely in terms of demographic and socioeconomic characteristics, minimum standards of reliability must be achieved in all HIS populations to test hypotheses that involve cross-population comparisons. Ware, Snyder, and Wright (1976) reported that data quality (e.g., reliability) tends to be poorer in more disadvantaged groups when standardized instruments are fielded. To ensure that reliability was adequate in the "worst" case, HIS staff examined the reliability of scales for groups differing in educational attainment. (The definition of these groups differed slightly for physical and mental health scale reliability studies; see Vols. II and III for details.)

Reproducibility. The concept of reproducibility—the degree to which a person's item responses can be predicted from knowledge of his total score—is closely related to reliability, and is thus an appropriate indicator of the reliability of Guttman scales. The coefficient of reproducibility (CR), discussed with respect to Guttman Scalogram Analysis above, defines a special case of internal-consistency reliability. The internal-consistency of a group of items represents the degree to which they measure the same construct. If CR is high, a Guttman scale is both reproducible and internally consistent. If CR is low, either the scale is not internally consistent, or is not cumulative, or is neither. As noted above, CR values of 0.90 or greater were accepted as evidence of the reliability and reproducibility of Guttman scales, consistent with Guttman (1944) and Edwards (1957).

Comments. Generally speaking, the higher the reliability of a measure the better, because high reliability indicates that more true score rather than random error is obtained. Reliability can be increased for a specific construct by devoting more items to its measurement. When measurement resources are limited, however, a tradeoff exists between achieving a more reliable score for one measure and measuring additional constructs. For purposes of group comparisons in the HIS, a minimum standard of 0.50 was set for reliability. Therefore, rather than using extra items to achieve very high reliability coefficients (above 0.90), they will be used to measure other constructs of importance or will be eliminated to reduce respondent burden. Because the HIS has heterogeneous populations, however, the strategy of eliminating items will not be used until adequate reliability is demonstrated in all sites. There is no way to know in advance that internal-consistency reliability will be too high. Very high test-retest coefficients may be undesirable, particularly when the period between administrations is long (e.g., the one-year interval between health questionnaires). High coefficients in this case indicate that the health construct being measured is very stable and may not change as a result of the experiment. (An unlikely exception would be the case in which test-retest coefficients were very high, the rank order of individuals in the score distribution remained the same, and a significant mean difference in scores occurred between administrations, such as would happen with measures of age over time.)

As Vols. II and III discuss in detail, several item groupings for physical and mental health scales were revised during empirical scaling studies of Dayton data. Furthermore, health status batteries were revised for use in health questionnaires fielded after Dayton enrollment, and new item groupings were hypothesized following scaling studies performed on Dayton data. Because revisions in item groupings were made on the basis of scaling criteria, particularly those associated with the Summated Ratings Method, that maximize the internal-consistency of scales, reliability estimates for these scales will be cross-validated before conclusions about reliability are drawn. To the extent that scale revisions capitalized on chance relationships among items in one site, reliability estimates for scales that were revised on the basis of empirical scaling studies may decrease on cross-validation. Although this is unlikely, in view of the generally high reliability estimates reported by other investigators for scales similar to those constructed in the HIS, the possibility must be tested during spot checks of scaling decisions using data from other HIS sites.

Studies of Validity

In addition to being reliable, the validity of a measure must be well understood before it can be used in testing hypotheses. Although reliability studies provide estimates of how much information (true score variance) is provided by a measure, validity studies are necessary to determine what should be inferred about the meaning of scores. Unless a measure is judged valid, the scores it yields cannot be interpreted with confidence for purposes of hypothesis testing, and the measure cannot be used to advance theory (e.g., by studying relationships between the measure and other variables of interest). In the case of HIS measures, issues of validity in relation to health status were addressed at three levels:

1. Does each measure yield information about health status?
2. Does each measure yield information about the specific health construct it was intended to assess rather than other health constructs?
3. To what extent is each measure a valid indicator of other variables (e.g., as a predictor of health and illness behavior)?

The first task in studying the validity of HIS measures was determining the appropriateness of using MHQ data to test hypotheses about the effects of insurance on health status. More specifically, measurement validation focused on how group differences, when observed, should be interpreted. Further, because the HIS data files will eventually become a national resource for secondary analyses, it was also important to provide as much information as practical about the validity of HIS measures so that others could judge their suitability for use in other studies and whether they represent improvements over similar measurement batteries.

Although the physical and mental health measures whose validity is discussed in the R-1987-HEW series were adapted from measures that had previously been fielded, available evidence pertinent to validity issues differed for each, and none of the original measures had been extensively validated. Furthermore, because of differences in the operational definitions of HIS measures, validation of some measures was more problematic than others. This problem was further complicated by gaps in theoretical and empirical analyses of relationships among health constructs, as well as between these constructs and others to which they should and should not relate if they in fact measure the aspect of health that was intended. The latter complication was particularly true for both mental and social health measures.

Several methods were used to evaluate validity; evaluation involved synthesizing information across these methods and making a judgment as to what construct(s) the measure reflected most. A successful evaluation will indicate both that the scale measures the one construct it was designed to measure and does not measure any other construct. The American Psychological Association's (1974) standards on how the validity of measures should be evaluated greatly influenced HIS validity studies. Three types of validity are identified by the Association: content validity, criterion-related validity, and construct validity.

Content Validity. Content validity refers to how well a measurement battery covers important aspects of the health dimension to be measured. HIS studies of content validity were designed to determine whether HIS measures were comprehensive with respect to the categories of health status selected for measurement. First, the batteries were judged to determine whether all three major dimensions of health status (in the HIS, physical, mental, and social) were represented. Second, each battery was judged to determine whether all relevant aspects of each dimension were adequately represented (e.g., whether anxiety, depression, positive well-being, and self-control constructs appeared in the mental health battery). These judgments were made considering the HIS goals of health status measurement, discussed at length in Sec. I. As a result, content validity was judged acceptable for HIS purposes in several cases where such a judgment might not have been made by investigators desiring more comprehensive measures.

Face validity, related to but distinct from content validity, refers to what an item appears to measure from its manifest content. To evaluate face validity, HIS staff reviewed the words used in each item to determine their relevance and adequacy as descriptors of the health construct to be measured. For example, did items including such words as "depressed," "sad," "tense," "relaxed," "cheerful," and "happy with life" describe psychological states that reflected mental health? Did activities such as light housework, running a block, requiring crutches to get around, and being confined to bed all day reflect physical health or lack thereof?

Face validity has been referred to as the "appearance of validity" and for this reason is not generally accepted as a basis for interpreting scores (American Psychological Association, 1974). Moreover, unlike criterion-related and construct validity, face and content validation are not empirically based. On these grounds, there seems to be a general prejudice against relying on evidence of face and content validity. Although no one validation approach is sufficient, face and content validation were regarded as a very appropriate first step in understanding the meaning of HIS health status measures. This opinion was based on several considerations. First, analyses of face and content validity are relatively easy to perform once the health concepts appropriate to the study have been identified. Second, examination of content reveals much about what the items measure and the meaning of responses to them. Particularly important, this examination can help avoid the problems of confounded definitions and misleading or contradictory labels. For example, a battery of items may be labeled as a measure of one health dimension but include items assessing several health dimensions; batteries of very similar content may receive very different labels, while those of dissimilar content may be given the same label. Third, problems revealed in empirical validation can generally be anticipated by careful face and content validation studies. Experience indicates no instances in which hypotheses about the meaning of general population health measures, such as those discussed in the literature reviews and developed for HIS use, that were formulated after thorough face and content validity studies have been contradicted by the results of empirical validation.

Criterion-Related Validity.[16] Criterion-related validity is assessed by using a person's score on one measure to predict his score on some other measure referred to as a "criterion." Strictly speaking, the term "criterion" is reserved for a previously validated measure of the same construct, which provides an estimate of the person's true score on the measure being validated. Evidence of criterion validity is often expressed as a correlation between two measures. If both are obtained during the same interval, the correlation is usually referred to as evidence of "concurrent" validity. If sufficient time elapses between the two measurements for change to occur, the correlation is usually termed evidence of "predictive" validity.

Criterion validation of HIS health status measures posed two problems. First, the validity of potential criterion measures was as much open to question as that of the HIS measures being validated. Second, many measures considered "criteria" in the literature (e.g., assessments by physicians in the case of functional status, ratings by friends and relatives in the case of social health) assessed constructs related to but different from the construct the HIS measure attempted to assess. Many of these coefficients are discussed in the R-1987-HEW series, although they are not termed evidence of "criterion validity." These coefficients provide useful information about HIS measures and allow comparison between HIS measures and those discussed in the literature; they also provide information to others interested in specific kinds of predictions (e.g., predicting use of health care services). Because well-validated measures that would provide true score estimates were not available for the health status constructs of interest to the HIS, criterion-related validity studies were not performed for HIS measures.

Construct Validity. Construct validity is assessed by examining the patterns of relationships between the measure being validated and measures of other variables theoretically related, or unrelated, to it. Validity is supported when the associations show the direction and magnitude of relationships hypothesized from theory. When exceptions to hypotheses are observed and theory is well founded, validity should be questioned. When there is reason to question both theory and measurement validity, drawing inferences about validity is more difficult.

Like criterion-related validation, construct validation usually relies on correla-

[16] See Anastasi (1968) and Cronbach (1970).

tion coefficients as evidence of hypothesized relationships. Unlike criterion-related validity studies, however, construct validity studies are multivariate. Because of the shortcomings of a single criterion variable or the absence of an agreed-upon criterion, a network of relationships is examined. Little can be learned about validity from the correlation between a measure and one other measure, unless the latter is an established criterion. Further, not even the correlation between a measure and a criterion provides a complete picture of validity. A single correlation says nothing about other variables with which a measure may correlate or about whether the measure is independent of other sources of influence. For these reasons, construct validation requires examination of many variables.

Another important attribute of construct validation is its reliance on a synthesis of findings from many independent studies. The in-depth understanding of the meaning of a score that is necessary for proper interpretation of results is not likely to occur following a single investigation. Therefore, synthesis of results across studies is necessary as validity issues are raised and resolved. Replication is also important to ensure that conclusions about validity of measurement can be generalized.

Reviews of the literature on measurement of physical, mental, and social health and general health perceptions were done in part to identify theoretical and empirical studies of such associations. Based on findings reported in the literature and on theoretical considerations, hypotheses regarding the strength and direction of relationships that might be expected were proposed. To the extent that relationships conformed to hypotheses, they supported both the construct validity of the measure and the theory underlying the relationships. In some cases, gaps in theoretical and empirical validity analyses reported in the literature, and conflicting results and interpretations (particularly for mental and social health measures), complicated HIS validity studies. When relationships of interest had not been previously tested, or when theory was not clear on the expected direction and magnitude of association, attempts at validation were hampered. In such cases, null hypotheses were proposed and tested to clarify measurement validity and the theory being studied.

All HIS health status measures were hypothesized to relate significantly to each other, based on the belief in a general health status construct that is common to all health status dimensions. Further, measures of different constructs within the same dimension (e.g., functional limitations and physical abilities measures within the physical health dimension) should be more highly related to each other than to measures of other dimensions (e.g., mental health). This hypothesis was based on the belief that constructs in the same dimension share variance because of a general health factor and because of the specific dimension they were hypothesized to measure. Similarly, high intercorrelations would be expected for measures of mental health constructs; anxiety and depression should be more highly related to each other than to measures of physical health constructs, because if these mental health measures are valid they will have both general and specific variance in common.

For some HIS health status measures, validity hypotheses were more complicated because the measures were intended to assess more than one dimension of health status. For example, measures of general health perceptions should be sensitive to differences in all three health dimensions, because people may consider physical, mental, and social health when rating their overall health.

In studying construct validity, it is also important to demonstrate that constructs other than health status (e.g., patient satisfaction, life changes) measured using the same method (a self-administered questionnaire in the case of the MHQ) are not highly related to HIS health status measures. To test this type of hypothesis (usually referred to as discriminant validity), measures of variables other than health status were included in the validity studies. Of course, such variables as life changes and patient satisfaction should be related to health status; however, if a measure of one health status construct is valid, it should be more related to other measures of the same health status construct than to measures of other constructs.

Statistics sensitive to both linear and nonlinear relationships were used in HIS studies of construct validity, and in many instances bivariate scatter plots were inspected visually to better understand relationships. Most relationships studied were roughly linear and were well represented by product-moment correlations; where curvilinearity was prominent or published studies showed a preference for nonparametric estimates of associations, gamma (Goodman and Kruskal, 1954) and eta coefficients (Thorndike, 1978) were also reported.

The hypotheses summarized above were tested by examining matrices of correlations between health status measures and measures of other variables included for validation purposes. Several different matrices were computed for these studies. For the physical and mental health scales, matrices of inter-scale correlations were computed to study associations among measures within each of these dimensions. To further explore relationships between these dimensions and the health-related variables to which they should be related, physical health scales were correlated with measures of physical exercise and mental health scales with measures of stress, recognition of mental problems, use of mental health care, and life satisfaction. These matrices were small enough that patterns of associations could be interpreted visually.

To study relationships among all HIS health status measures and selected health-related variables, a large symmetrical matrix of correlations among all health status measures and selected validity variables was computed.[17] This matrix contained 28 variables and 378 nonredundant correlations. For such a matrix, "eyeball" inspection would have been very time consuming, complicated, and possibly misleading. For these reasons, the information contained in this validity matrix was summarized using factor analysis.

Factor analysis is an empirical method for determining the nature and number of dimensions (in this case, dimensions of health status) that account for correlations among measures. These dimensions or factors are extracted from the matrix in the order of their importance, that is, how much of the variance they account for in all the measures. For the matrix of correlations among all HIS health status measures, more than one orthogonal (statistically uncorrelated) factor was hypothesized because more than one dimension was represented by the variables. In fact, because measures in the matrix were hypothesized to represent physical, mental, and social dimensions of health status, three important orthogonal factors were hypothesized.

The factor analysis was conducted in three steps. First, factors were extracted

[17] Studies of the physical and mental health scales are reported in the appropriate individual volumes; to minimize redundancy in presenting construct validity findings, multivariate studies involving all HIS health status constructs are discussed only in Vol. VI.

from the matrix. Correlations between the measures and the first factor in the unrotated matrix indicated how much each measure was explained by a general common dimension. Thus, the first HIS hypothesis, that all measures of health status would have some variance in common, was tested by examining their correlations with the first unrotated factor.

In the second step, factors were rotated to orthogonal structure to facilitate their interpretation.[18] Each factor was interpreted by inspecting its correlations with health status measures and validity variables included in the matrix. For example, a factor that correlated highly with many mental health measures and did not correlate highly with other measures would be interpreted as a dimension of mental health.

Finally, once each factor was interpreted, it could be used as a criterion to test the validity of each measure in question. This is a major advantage of the factor analytic approach to construct validation; it can be used when no validity criteria exist. However, because the factors are used as criteria against which to interpret the validity of each measure, only those factors that are important and can be interpreted unequivocally should be retained. Interpreting measures in terms of factors that are themselves in question is inherently ambiguous.

Factor analysis does have certain limitations. First, interpreting factors can be subjective. Objectivity in the HIS analyses was maximized by including several measures of each major dimension, and using different scaling methods and different definitions of constructs within the same dimension. When factor analytic results are very clear in a well-defined matrix (i.e., with multiple indicators of each underlying dimension), subjectivity of interpretation is minimized. Second, correlations between measures and factors define the relationships at a point in time. Each correlation reflects both cause and effect relationships between a measure and a factor and the influence of other variables that affect both measure and factor. Therefore, interpretations and especially causal inferences must be made very cautiously. Finally, the results of factor analysis (or any multivariate analysis) are only as good as the information contained in the measures. For HIS analyses, data quality tended to be very high. Reliability coefficients were moderate to high (greater than 0.70 up to 0.99) in Dayton, and considerable information about the measures was available from previous research and from the scaling studies that preceded factor analysis. Thus, in this instance, factor analysis was justified.

Power Analyses

An important feature of any health status measure is its ability to detect differences, if they exist, among groups of individuals. In the HIS, the differences that must be detected are those among groups enrolled in different insurance plans. Power is defined as the probability of correctly rejecting the null hypothesis of no group differences when a particular contrary hypothesis is true. In the HIS, 1 − power = beta, which is the probability of making a Type II error, or alternatively the probability of accepting a null hypothesis when it is false.

Estimates of the power of a measure can be expressed in several ways. All take into consideration hypotheses to be tested, method of analysis, experimental design, points in time at which the information on health status is obtained, and

[18] The unrotated matrix and criteria used during factor rotation were discussed above under "Factor Analysis" in the section on "Construction and Evaluation of Multi-Item Scales."

variability of score distributions on the measures. Estimates of the power of HIS health status measures were expressed in terms of differences in health status or effect sizes between two groups or insurance plans that one wishes to detect. Specifically, they were expressed in terms of the percentage difference in mean scale scores that can be detected across plans, and in terms of differences in scores predicted by other relevant variables. Effect sizes for physical health scales were expressed in terms of differences in age and those for mental health scales in units of stressful life changes.

HIS power analyses (Vol. VII) provided information on the magnitude of group differences in health status that can be detected under the conditions of the HIS. Estimates were calculated assuming the HIS sample size of some 3900 adults grouped into a free plan (one-third) and all nonfree plans (two-thirds), availability of pretests and covariates that would yield intertemporal correlations between measures similar to those reported in the literature, linear regression methods, and conventional error rates (probabilities of Type I error of 0.05 and Type II error of 0.10). (A Type I error refers to rejecting a null hypothesis when it is true.) Power analyses were carried out only for health status scales constructed in Dayton (i.e., physical and mental health); Dayton data provided estimates of mean scores and standard deviations on these scales for the entire HIS sample.

For several reasons, the precision of HIS measures could be either better or worse than indicated. In the power calculations performed on Dayton data, an attempt was made to be conservative. First, as discussed in other volumes of this series, revised versions of the Dayton measures were used at enrollment in all sites after Dayton and on all repeat administrations in all sites. This difference should be kept in mind when interpreting the power findings reported in Vol. VIII, because the revisions should yield measures of higher power. Second, certain analytic refinements that would affect power, such as repeated measurement, intra-family correlation, and stratification, were not taken into account. Third, changes in the sample or in data quality (e.g., missing data caused by death or other attrition from the sample) were not considered. Fourth, linear regression methods of estimation and hypothesis testing were presumed, although more efficient techniques may be available. Finally, some leeway should be allowed for the surprises that inevitably accompany careful analysis of complex data.

REFERENCES

American Psychological Association, *Standard for Educational & Psychological Tests,* American Psychological Association, Washington, D.C., 1974.

Anastasi, A., *Psychological Testing,* 3d ed., Macmillan Publishing Co., Inc., New York, 1968.

Armor, D. J., "Theta Reliability and Factor Scaling," in H. L. Costner (ed)., *Sociological Methodology, 1973-1974,* Jossey-Bass, San Francisco, 1974.

Bradburn, N. M., *The Structure of Psychological Well-Being,* Aldine Publishing Company, Chicago, 1969.

Brook, R. H., et al., *Conceptualization and Measurement of Physiologic Health for Adults in the Health Insurance Study,* The Rand Corporation, R-2262-HEW, forthcoming.

Campbell, D. T., and D. W. Fiske, "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix," *Psychological Bulletin* 56: 81-105, *1959.*

Cattell, R. B. (ed.), *Handbook of Multivariate Experimental Psychology,* Rand McNally & Company, Chicago, 1966.

Clasquin, L. A., and M. E. Brown, *Rules of Operation for the Rand Health Insurance Study,* The Rand Corporation, R-1602-HEW, May 1977.

Comrey, A. L., "Tandem Criteria for Analytic Rotation in Factor Analysis," *Psychometrika* 32:143-154, 1967.

Comrey, A. L., *A First Course in Factor Analysis,* Academic Press Inc., New York, 1973.

Cronbach, L. J., "Coefficient Alpha and the Internal Structure of Tests," *Psychometrika* 16:297-334, 1951.

Cronbach, L. J., *Essentials of Psychological Testing,* 3d ed., Harper & Row, Publishers, New York, 1970.

Edwards, A. L., *Techniques of Attitude Scale Construction,* Appleton-Century-Crofts Inc., New York, 1957.

Eisen, M., et al., *Conceptualization and Measurement of Health for Children in the Health Insurance Study,* The Rand Corporation, R-2313-HEW, May 1980.

Fine, M., "Interrelationships among Mobility, Health and Attitudinal Variables in an Urban Elderly Population," *Human Relations* 28: 451-474, 1975.

Fishbein, M. (ed.), *Readings in Attitude Theory and Measurement,* John Wiley & Sons Inc., New York, 1967.

Fiske, D. W., "Some Hypotheses Concerning Test Adequacy," *Educational and Psychological Measurement* 26:69-88, 1966.

Gilson, B. S., et al., *Further Tests and Revisions of the Sickness Impact Profile 1974-75,* Department of Health Services, School of Public Health and Community Medicine, Seattle, Wash., 1975.

Goodman, L. A., and W. H. Kruskal, "Measures of Association for Cross Classifications," *Journal of the American Statistical Association* 49:732-764, 1954.

Greenblatt, H. N., *Measurement of Social Well-Being in a General Population Survey,* Human Population Laboratory, California State Department of Health, Berkeley, 1975.

Guertin, W. H., and J. P. Bailey, Jr., *Introduction to Modern Factor Analysis,* Edwards Brothers, Inc., Ann Arbor, Mich., 1970.

Guttman, L. A., "A Basis for Scaling Qualitative Data," *American Sociological Review* 9:139-150, 1944.

Helmstadter, G. C., *Principles of Psychological Measurement,* Appleton-Century-Crofts Inc., New York, 1964.

Howard, K. I., and G. G. Forehand, "A Method for Correcting Item-Total Correlations for the Effect of Relevant Item Inclusion," *Educational and Psychological Measurement* 22:731-735, 1962.

Jackson, D. N., and S. Messick (eds.), *Problems in Human Assessment,* McGraw-Hill Book Company, New York, 1967.

Jeffers, F. C., and C. R. Nichols, "The Relationship of Activities and Attitudes to Physical Well-Being in Older People," *Journal of Gerontology* 16:67-70, 1961.

Kane, R. A., R. H. Brook, and R. L. Kane, *Conceptualization and Measurement of Health Habits for Adults in the Health Insurance Study: Vol. III, Alcohol Consumption,* The Rand Corporation, R-2374/3-HEW, forthcoming.

Kerlinger, F. N., and E. J. Pedhazur, *Multiple Regression in Behavioral Research,* Holt, Rinehart, & Winston Inc., New York, 1973.

Klemmack, D. L., J. N. Edwards, and J. R. Carlson, "Measures of Well-Being: An Empirical and Critical Assessment," *Journal of Health and Social Behavior* 15:267-270, 1974.

Likert, R., "A Technique for the Measurement of Attitudes," *Archives of Psychology* 140:1-55, 1932.

Morris, C., "A Finite Selection Model for Experimental Design of the Health Insurance Study," *Journal of Econometrics* 11:43-61, 1979.

Newhouse, J. P., "A Design for a Health Insurance Experiment," *Inquiry* 11:5-27, 1974.

Newhouse, J. P., "Insurance Benefits, Out-of-Pocket Payments, and the Demand for Medical Care," *Health and Medical Care Services Review* 1:1,3-15, July-August 1978.

Newhouse, J. P., C. E. Phelps, and W. Schwartz, "Policy Options and the Impact of National Health Insurance," *New England Journal of Medicine* 290:1345-1359, 1974.

Nie, H. H., et al., *SPSS: Statistical Package for the Social Sciences,* 2d ed., McGraw-Hill Book Company, New York, 1975.

Nunnally, J. C., *Psychometric Theory,* McGraw-Hill Book Company, New York, 1967.

Palmore, E., and C. Luikart, "Health and Social Factors Related to Life Satisfaction," *Journal of Health and Social Behavior* 13:68-80, 1972.

Schooler, C., "A Note of Extreme Caution on the Use of Guttman Scales," in G. M. Maranell (ed.), *Scaling: A Sourcebook for Behavioral Scientists,* Aldine Publishing Company, Chicago, 1974, pp. 223-230.

Smith, L. H., et al., *The Health Insurance Study Screening Examination Procedures Manual,* The Rand Corporation, R-2101-HEW, September 1978.

Social Psychiatry Research Unit, "The Psychiatric Epidemiology Research Interview: A Report on Twenty-Two Scales, Appendix I to the Measurement of Psychopathology in the Community," Columbia University Press, New York, February 1977 (mimeographed).

Stewart, A. L., R. H. Brook, and R. L. Kane, *Conceptualization and Measurement of Health Habits for Adults in the Health Insurance Study: Vol. I, Smoking,* The Rand Corporation, R-2374/1-HEW, June 1979.

Stewart, A. L., R. L. Kane, and R. H. Brook, *Conceptualization and Measurement of Health Habits for Adults in the Health Insurance Study: Vol. II, Overweight,* The Rand Corporation, R-2374/2-HEW, 1980.

Stewart, A. L., R. L. Kane, and R. H. Brook, *Conceptualization and Measurement of Health Habits for Adults in the Health Insurance Study: Vol. IV, Exercise,* The Rand Corporation, R-2374/4-HEW, forthcoming.

Thorndike, R. M., *Correlational Procedures for Research,* Gardner Press, Inc., New York, 1978.

Thurstone, L. L., and E. J. Chave, *The Measurement of Attitude,* University of Chicago Press, Chicago, 1929.

Tyler, T. A., and D. W. Fiske, "Homogeneity Indices and Text Length," *Educational and Psychological Measurement* 28:767-777, 1968.

Ware, J. E., Jr., W. G. Miller, and M. K. Snyder, *Comparison of Factor Analytic Methods in the Development of Health-Related Indexes from Questionnaire Data,* Publication No. PB-239-517/AS, National Technical Information Service, Springfield, Va., 1973.

Ware, J. E., Jr., M. K. Snyder, and W. R. Wright, *Development and Validation of Scales To Measure Patient Satisfaction with Health Care Services: Volume I of a Final Report, Part B: Results Regarding Scales Constructed from the Patient Satisfaction Questionnaire and Measures of Other Health Care Perceptions,* NCHSR Report No. 7918, NTIS Publication No. PB-288-330, National Technical Information Service, Springfield, Va., 1976.

White, B. W., and E. Saltz, "The Measurement of Reproducibility," in D. N. Jackson and S. Messick (eds.), *Problems in Human Assessment,* McGraw-Hill Book Company, New York, 1967.

World Health Organization, *Constitution,* in *Basic Documents,* World Health Organization, Geneva, 1948.