Volume VI
Report No. 63
November, 1966

Psychiatric Research
and the
Assessment of Change

Formulated by the
Committee on Research

Group for the Advancement of Psychiatry
Table of Contents

INTRODUCTION ........................................... 357

1. DETERMINANTS OF RESEARCH CHOICES .......... 360
   Personal Value Orientations and Commitments 361
   Sociocultural Influences .......................... 363
   The State of Development of the Science ....... 365
   Scientific Models and Scientific Identifications 366
   The Philosophies of Research ...................... 372

2. CONCEPTS OF CHANGE ............................... 380
   Context of Change ................................ 380
   Kinds of Change .................................. 383
   Dimensions of Change .............................. 387

3. FACTORS INFLUENCING THE ASSESSMENT OF CHANGE 392
   The Subject ...................................... 392
   The Observer .................................... 393
   The Concepts .................................... 394
   Time and Sequence ................................ 395
   Space and Field ................................. 397

4. THE MEASUREMENT OF CHANGE ...................... 399
   Theory of Measurement ............................. 399
   Problems of Measurement: The Phenomenon as the Variable ................................. 414
   The Setting and the Subject as Variables .... 426
   The Investigator and the Instrument as Variables ............................................. 434

5. PROBLEMS OF DESIGN AND OF CONFIRMATION .... 453
   Research Strategies ................................ 453
   Replication ....................................... 456
   The Use of Controls ................................. 468

6. SUMMARY ............................................. 470

REFERENCES ........................................... 472

This report is the fourth in a series of Reports and Symposia that will comprise Volume VI. For a list of other GAP publications on topics related to the subject of this report, please see page 481.
STATEMENT OF PURPOSE

The GROUP FOR THE ADVANCEMENT OF PSYCHIATRY has a membership of approximately 185 psychiatrists, organized in the form of a number of working committees which direct their efforts toward the study of various aspects of psychiatry and toward the application of this knowledge to the fields of mental health and human relations. Collaboration with specialists in other disciplines has been and is one of GAP's working principles. Since the formation of GAP in 1946 its members have worked closely with such other specialists as anthropologists, biologists, economists, statisticians, educators, lawyers, nurses, psychologists, sociologists, social workers, and experts in mass communication, philosophy, and semantics. GAP envisions a continuing program of work according to the following aims:

1. To collect and appraise significant data in the field of psychiatry, mental health, and human relations;
2. To re-evaluate old concepts and to develop and test new ones;
3. To apply the knowledge thus obtained for the promotion of mental health and good human relations.

GAP is an independent group and its reports represent the composite findings and opinions of its members only, guided by its many consultants.

PSYCHIATRIC RESEARCH AND THE ASSESSMENT OF CHANGE WAS FORMULATED BY THE COMMITTEE ON RESEARCH. The members of this Committee as well as all other Committees are listed below.

* The Committee is deeply indebted to its long-time consultants who, throughout, shared fully in its tasks of formulation and execution of this report, with, in addition, special concern for their own areas of expertise as indicated. They are: Erwin A. Haggard, Ph.D., Professor of Psychology, Department of Psychiatry, University of Illinois College of Medicine (problems in measurement and in qualification); Jerome Richfield, Ph.D., Professor and Chairman, Department of Philosophy, San Fernando Valley State College, California (problems in philosophy, logic, and values). During the first half of its work, the Committee benefited markedly from the synthesizing labors of another consultant, Richard M. Williams, Ph.D., Special Assistant to the Director, NIMH. In addition, three GAP Fellows, during their term of service with the Committee, contributed to the thinking, the writing, and the rewriting of this report: Everett Dulll (Albert Einstein College of Medicine), Harold Spira (Johns Hopkins University School of Medicine), and Harris S. Goldstein (Sheppard and Enoch Pratt Hospital). John Tucey, Ph.D., Professor of Mathematics, Princeton University, met with the Committee once as consultant in the field of measurement theory. The consultants are all represented in the product by the same variegated and manifold ways as are the Committee members themselves.

COMMITTEE ON RESEARCH
Robert Wallis Stein, San Francisco, Calif.
G. E. Bitting, Boston
Stevan Eide, Belmont, Mass.
Evelin F. Gildea, St. Louis
Louis A. Gottschalk, Cincinnati
Shepperd G. Kallman, Chicago
Gerald L. Kierman, New Haven
Morris A. Liptzin, Cherry Chase
Ralph R. Nottman, Brookline, Mass.
Franz Reichmann, Brooklyn
Richard E. Benneker, Los Angeles
Alfred H. Stalin, Boston

COMMITTEE ON ADOLESCENCE
Calvin Settle, Ardmore, Pa., Ch.
Walter J. Gadsbille, Denver, Colo.
Mary O. Nethal Hawkins, New York
Joseph J. Michaels, Belmont, Mass.
Joseph P. Noshpitz, Washington, D.C.
Vivian Raskoff, Montreal
Harry Wermer, Chiswick, Mass.

COMMITTEE ON AGING
Jack Weingberg, Chicago, Ill.
Robert N. Butler, Washington, D.C.
Avish K. Goldfarb, New York
Lawrence Greensligh, Los Angeles
Maurice E. Lindner, Philadelphia
Prescott Thompson, Topanga
Montague Ullmann, Brooklyn

COMMITTEE ON CHILD PSYCHIATRY
Susanne T. van Amerongen, Ecusta, Ch.
R. James Anthony, St. Louis
H. Donald Burton, New York
John F. Ewerhard, Chicago
William S. Leegford, New York
Dame O. Foy, Durner, Colo.
Ecole B. Wold, New York

COMMITTEE ON COLLEGE STUDENT
Harriette E. Eddy, New York, Ch.
Robert L. Arnowit, New Haven
Alfred Flammstein, Chicago
Alan Frank, Boulder, Colo.
Erie Silber, Washington, D.C.
COMMITTEE ON MENTAL HOSPITAL SERVICES
Lucy D. Ozarlin, Bethesda, Md.
Walter E. Barton, Washington, D.C.
Merrill Eaton, Omaha
James B. Funkhouser, Richmond, Va.
Francis J. O'Neill, Central Islip, N.Y.
Lee S. Swayne, N. Little Rock, Ark.
Jack A. Wodford, Pittsburgh

COMMITTEE ON MENTAL REHABILITATION
Leo Madow, Philadelphia, Pa.
Howard V. Bair, Parsons, Kan.
Peter W. Brown, Pownal, Me.
Sue M. Finch, Ann Arbor
George Tuzman, Los Angeles
Warren T. Vaughn, Jr., San Mateo
Thomas G. Webster, Bethesda
Henry H. Work, Los Angeles

COMMITTEE ON PREVENTIVE PSYCHIATRY
Leonard J. Dubin, Washington, D.C., Md.
Gerald Crapian, Boston
Jules V. Coleman, New Havent
Stephen Plock, New Haven
Albert J. Glass, Oldham City
Mary E. Meroor, New York
Marvin M. Perkins, New York
Harold M. Visotsky, Chicago
Stanley F. Yolles, Bethesda

COMMITTEE ON PSYCHIATRY IN INDUSTRY
Harry E. Wagenheim, Philadelphia, Pa.
Spencer Byllesby, Houston
Thomas L. Bromnick, Rochester, Minn.
Matthew Brody, Brooklyn
Alan A. McLean, New York
Kenneth J. Mundon, Memphis
Oubban C. Taylor, Montreal

COMMITTEE ON PSYCHIATRY AND LAW
Lena L. Usdin, New Orleans, La.
Zigmund M. Lembernitz, Washington, D.C.
Edward T. Auer, St. Louis

John Donnelly, Hartford
Jay Katz, New Haven
Andrew S. Watson, Ann Arbor

COMMITTEE ON PSYCHIATRY AND RELIGION
Mortimer Osnow, New York, N.Y.
Sidney Furst, New York
John W. Higgin, St. Louis
Stanley A. Leavy, New Haven
Earl A. Loomis, Jr., New York
Albert J. Lubin, Woodside, Calif.

COMMITTEE ON PSYCHIATRY AND SOCIAL WORK
Edward C. Franz, Louisville, Ky.
C. Knight Aldrich, Chicago
Maurice R. Friend, New York
John MacLeod, Cincinnati
John Newall, Boston
Eleanor A. Steele, Denver
Edward M. Weinshel, San Francisco

COMMITTEE ON PSYCHOPATHOLOGY
Oswald C. Martin, Roanoke, Va.
Wagner H. Bridger, New York
Neil B. Burch, Houston
Daniel X. Freedman, New Haven
Paul E. Houston, Iowa City
P. Herbert Leidner, Palo Alto, Calif.
George Ruff, Philadelphia
Charles Shapen, Philadelphia
Albert J. Silverman, New Brunswick, N.J.

COMMITTEE ON PSYCHOPATHOLOGY
Milton Grossblatt, Boroa, N.Y.
Wagner H. Bridger, New York
Neil B. Burch, Houston
Daniel X. Freedman, New Haven
Paul E. Houston, Iowa City
P. Herbert Leidner, Palo Alto, Calif.
George Ruff, Philadelphia
Charles Shapen, Philadelphia
Albert J. Silverman, New Brunswick, N.J.

COMMITTEE ON PUBLIC EDUCATION
Peter A. Martin, Detroit, Mich.
Leo H. Bartholmez, Baltimore
H. Waldo Bird, St. Louis
Lloyd C. Eaton, Nashville
Dana L. Farnsworth, Cambridge
John P. Lambert, Kalamazoo, Mich.
Michael Michael Baisman, Charleston
Mabel Ross, New York
Mathew Ross, Chestnut Hill, Mass.
Julius Schruder, Washington, D.C.
Kent A. Zimmerman, Berkeley

COMMITTEE ON SOCIAL ISSUES
Victor W. Bernard, New York
Robert Coles, Cambridge
Leslie Griswold, Boston
Joel S. Handler, Chicago
Ferdin Bell, Rome, New York
Judd Macmen, Los Angeles
Roy W. Meagher, Topeka
Charles A. Pinderhughes, Boston

COMMITTEE ON THERAPEUTIC CARE
Benjamin Simeon, Boston, Mass.
Marvin L. Adlen, Chevy Chase, Md.
Issi L. W. Charnley, Ottawa
Thomas E. Curtis, Chapel Hill
Robert W. Gilson, Tewson, Md.
Harold A. Greenberg, Bethesda
Henry R. Grubban, Boston
Bernard H. Hall, Topeka
Leiser H. Levy, Chicago
Malcolm Sabahia, Chicago
Robert H. Switzer, Topeka

COMMITTEE ON THERAPY
Peter H. Kapp, Boston, Mass.
Henry W. Brown, Pittsburgh
Harvey H. Corman, New York
Eugene Meyer, Baltimore
William G. O'Keefe, Chicago
Albert E. Scholten, St. Louis
Harley C. Shands, Brooklyn
Lucie E. Tower, Chicago

CONTRIBUTING MEMBERS
Nathan W. Ackerman, New York
Charles C. Alden, Jr., Williamsbridge, N.Y.
Kenneth E. Appel, Philadelphia
M. Royden C. Astley, Edinburgh
Charlotte Babcock, Pittsburgh
Harley D. Baer, Florissant
Grace Balcer, New York
Benjamin H. Boley, New York
Bernard Randle, Boston
Alfred Paul Bay, Topeka
Anna R. Baugham, Chicago
A. E. Bennett, Berkeley
Robert E. Bennett, Princeton, N.J.
Ivan C. Benzine, Coral Gables, Fla.
Sidney Berman, Washington, D.C.
Edward G. Billings, Denver
Carl A. L. Binger, Cambridge
Daniel Blaine, Philadelphia
Wilfred Bloomberg, Harvard
C. H. Herdman, New York
Eugene Brody, Baltimore
Edward V. Butts, Durham
Dale C. Cameron, Washington, D.C.
Norman Cameron, New Haven
Ralph T. Collins, Rochester
Frank J. Curran, New York
Bernard L. Diamond, Berkeley
Franklin G. Eubanks, Denver
Joel Elkes, Baltimore
O. Spurgeon English, Nauberth, Pa.
Jack R. Ewalt, Boston
Jerome D. Frank, Baltimore
Lawrence Z. Freedman, Chicago
Thomas M. French, Chicago
Moses M. Frohlich, Ann Arbor
Dallas H. Fulkerson, Houston
George E. Gardner, Boston
J. S. Gottlieb, Detroit
Maurice H. Greenhill, Scarsdale, N.Y.
John H. Greist, Indianapolis
Roy R. Green, Chicago
Ernest M. Greenberg, New York
Manfred S. Guttmacher, Baltimore
David A. Hamburg, Palo Alto, Calif.
Frederick N. Hanson, New York
Herbert I. Harris, Cambridge
Leonard E. Hefler, Ann Arbor
J. Corpus Hesseburg, Topeka
Edward J. Hornick, New York
Roger William Howell, Ann Arbor
Joseph Hughes, Philadelphia
Robert W. Hyde, Boston
Lucien Jessner, Washington, D.C.
Irwin M. Jessenly, Phoenix
Marion B. Kowalsky, New York
Edward J. Kollan, Los Angeles
Othilda Krug, Cincinnati
Lawrence S. Kubie, Sparks, Md.
Paul V. Lenkai, Baltimore
Maurice Levine, Cincinnati
David M. Levy, New York
Harold J. Lief, New Orleans
Robert J. Litton, Woodbridge, Conn.
Erich Lindemann, Palo Alto, Calif.
Hymans S. Lippman, St. Paul
Reginald S. Louie, Washington, D.C.
Alfred G. Ludwig, Boston
LeRoy M. A. Maier, Philadelphia
Sydney G. Margolin, Denver
Helene V. McCall, Chicago
Karl Menninger, Topeka
Arthur A. Miller, Chicago
James G. Miller, Ann Arbor
Angel N. Miranda, Hato Rey, P.R.
Russell E. Monroe, Baltimore
Don P. Morris, Dallas
Peter B. Neubauer, New York
Rudolph G. Novick, Des Plaines, Ill.
Humphry F. Osmond, Princeton, N.J.
Eveleen N. Rexford, Cambridge, Mass.
Lewis L. Robbins, Glen Oaks, N.Y.
Milton Rosenbaum, New York
W. Donald Ross, Cincinnati
William M. Shanahan, Denver
Benjamin M. Spock, Cleveland
Edward Stinbrook, Los Angeles
Brandt F. Stebbins, Denver
Rutherford B. Stevens, New York
Lloyd J. Thompson, Chapel Hill
Harvey J. Tompkins, New York
Arthur V. Valenstein, Cambridge
Helen Stocken Wagenheim, Philadelphia
Cecil L. Whitman, Omaha
David Wright, Providence

LIFE MEMBERS
S. Spafford Ackely, Louisville
Earl D. Boud, Philadelphia
John R. Bree, London, England
Bruce Robinson, Calver, Ind.
Francis H. Sieper, Augusta, Me.

OFFICERS
President
Robert S. Garber
Carrer Clinic, Belle Mead, N.J. 08502

Vice President
Herbert C. Modlin
Box 829, Topeka, Kans. 66601

Secretary-Treasurer
Malcolm J. Farrell
Box C, Waverley, Mass. 02178

Asst. Secretary-Treasurer
Paul B. Hinton
500 Newton Road, Iowa City, Iowa 52241

PUBLICATIONS BOARD
Milton Greenblatt, Ch.
Boston State Hospital
591 Morton Street
Boston, Mass. 02124

Maurice R. Friend
Louis A. Gottschalk
Bernard H. Hall
Zigmund M. Levenson
John Nolan
Perry Ostemberg
INTRODUCTION

Despite the common feeling that “everybody knows what change is, and everyone is assessing it all the time,” the topic of the assessment of change is complex, embracing many of the knottiest conceptual problems and practical difficulties of our science, indeed of all science. Though the need for this report may not be at all peculiar to psychiatry, its preparation has highlighted a number of the major theoretical and practical issues in contemporary psychiatric research.

Research workers in related laboratory sciences, such as physiological psychology, or in related clinical disciplines, such as neurology, seldom express the same urgency as do psychiatrists to reflect upon the conceptual or the methodological foundations of their observations and their measurements. Psychiatrists, when they do substantive research, seem to need to keep one eye cocked on the soundness of their method, its relevance to their goal, and its adequacy as a source of reliable and valid substantive data much more than do workers in these related disciplines. These other fields have a more substantial methodological deposit; their techniques and their criteria rest upon a much larger—and more stable—substratum of agreed-upon concepts.

By comparison, psychiatric research has a shorter history and it differs, at least quantitatively, from these other more laboratory-based medical and behavioral sciences in a number of major respects:

1. The behavioral phenomena with which psychiatry deals are more subtle and complex, and often appear more elusive.

2. The investigative tradition is newer and the investigative procedures are less extensively developed.

3. Psychiatric research is often complicated by the additional role of therapy; that is, the psychiatrist, as therapist, tries to produce change, not just to measure it (this distinction is shared, of course, with other medical—cf. healing—disciplines).

4. Whether as therapist or not, the psychiatrist is more often actively involved in the process being investigated, he is himself more the instrument of the research, and his participation more decisively influences the phenomena (this is more true of psychiatry than of other medical disciplines).

5. Value judgments and semantic conventions are more often an issue in psychiatric research and psychiatric communication. There is less agreement on vocabulary, with consequent greater importance of linguistic reform and even innovation.

While these distinctions make the tasks in psychiatric research more difficult, they do not, in our opinion, render the problems insurmountable; rather, they offer unusual opportunities for critical self-appraisal and for imaginative new approaches. In dealing with these problems, the psychiatric research worker will in part be aided by the newer viewpoints in the philosophy of science and the newer concepts and techniques in mathematics, biology, and statistics. At the same time the lack of investigative models fully appropriate to the special circumstances of psychiatric phenomena cannot be met by the direct assimilation of models and techniques devised for the problems of other—even closely related—fields. Out of the creative encounter between the newer and more sophisticated trends in both psychiatry and its range of sister disciplines come both opportunities and dilemmas, toward the clarification of which this report seeks to contribute.

We have tried in this report to clarify some of these problems and issues as they are manifest in the assessment of change. Our main focus is on research in psychotherapy and other psychiatric
treatments, but we have also considered these problems as they arise in research in such areas of investigation as psychopharmacology, personality development and dynamics, and psychosomatic medicine. Since the assessment of change is a major daily concern of psychiatric practitioners, teachers, and administrators, as well as of researchers, we hope that this report will prove useful and provocative to all the diverse areas of psychiatry.

1

DETERMINANTS OF RESEARCH CHOICES

The assessment of change in psychiatric research involves many different factors that together lead the investigator to choose one frame of reference and perspective rather than another from which to examine change. With so many phenomena and variables available for scientific observation and from such varying points of view, it is pertinent to ask on what grounds investigators make decisions as to the facets of change on which attention should be focused. The pursuit of answers to such questions is often a fruitful step in determining some of the reasons for contradictory or disparate findings reported by two or more groups of qualified investigators examining the same phenomena.

The intent of this section on the considerations that lead to research choices is to suggest the range of human factors and of value orientations that, together with the scientific interest, determine the choice of what is relevant and salient in the phenomena of the field and in their research study. The nonscientific factors cover a wide spectrum of social, cultural, philosophical, and idiosyncratic value dispositions. They operate in the prosecution of scientific research at every step along the way: (1) in the choice of the conceptual framework within which the research will proceed; (2) in the selection of the area and the problem for study; (3) in the formulation of the research hypotheses and the specification of the governing assumptions; (4) in the formulation of the research design, including the specific choice of variables, of instruments, and of methods; (5) in the manner of making observations and gathering data; (6) in the methods of data reduction and analysis; (7) in the interpretation of the research.
findings; and (8) in the presentation of the findings to the scientific community. Actually, this entire sequence of choices is reflected in the investigator’s original protocol, in its statement of aims, of methods, and of values.

We do not regard these human value considerations in relation to problems of scientific choices only or even primarily as biases that lead to uncontrolled errors of observation and hence to incorrect conclusions, although under some circumstances they may operate in just this way. Rather, we are stating that all science is a human activity and that these same value considerations are the very forces that guide us in our determination of what it is significant to pursue in research; and that they are, in fact, the causes of our doing research at all. That is, proclivities and preconceptions are methodologically necessary to the conduct of all research; the problem is to be guided by fruitful ones and to avoid the possible pitfalls in even the most fruitful ones.

**Personal Value Orientations and Commitments**

The changes an investigator plans to observe are determined not only by his convictions regarding what variables are pertinent to the phenomena under study but also by the kinds of changes that personally interest him most. These two sets of determinants do not necessarily have all components in common.

The idea that it is more productive to collect data that can be quantified or treated mathematically than data not lending themselves to such measurement is a value orientation that can lead an investigator to examine only certain (quantifiable) kinds of change. In contrast, an investigator who feels the rigor and logic of mathematics to be inimical to the understanding of the most important personal dimensions of psychological phenomena may select for study variables that are not readily conceptualized in any conventional mathematical frame of reference. Possibly neither investigator will pay much critical attention to the other investigator’s focus of attention.

Some investigators are committed to the viewpoint that research is most fruitful when one seeks generalizations that apply broadly across all levels of organization, from the molecular through the organismic, from the social through the cosmic. An example of this strategic position is General Systems Research. Other investigators concentrate on variables appearing uniquely at one level of organization and that do not seem relevant to the phenomena occurring at other levels. Investigators of the latter type argue that what is lost in generalizing power is more than compensated for by discoveries of high pertinence and applicability—for example, at the human organismic level that is their central concern. Such differing viewpoints affect not only the variables chosen for examination but also the overall style of the research.

Another aspect of value commitment that plays an (often unrecognized) part in an investigator’s style of research is his overall pattern of personal value orientations, which at first sight seem well outside the range of relevance to his particular subject of scientific inquiry. Traditionally, such value orientations—whether political, aesthetic, religious, monetary, recreational, athletic, literary, or other—have been considered completely separate sets of mental operations, without influence on the nature of scientific observation. Yet the most casual consideration persuades us that how liberal a political position one espouses may importantly affect the way in which one sees the degree of possible environmental molding upon the hereditary givens in determining the final IQ score (theoretically a purely empirical scientific issue).

Psychoanalysis has pioneered in the elaborate study of the effect of the analyst’s extra-analytic interests and preferences on the process and outcome of analysis. Indeed, his self-analysis of such dispositions is a prescribed scientific obligation of the psychoanalyst conscientiously performing his role in the psychoanalytic procedure. Studies of this kind, although neither systematic nor rigorously controlled, have provided evidence that, in some instances, the political views—even the recreational preferences—of the psychoanalytic investigator may affect his perceptions of
and reactions to his analysand. This is not to speak of questions of clinical style, which importantly influence the way in which the therapeutic process unfolds, the climate within which it unfolds, and therefore the route to the desired clinical goal. Such questions of style, although linked to questions of value orientation, are by no means identical with them.

These many related questions invite more concentrated and systematic investigations in this whole area of the interrelationship of general value orientations, attitudes, predilections, and beliefs with the particular problems chosen for scientific inquiry and the kinds of variables chosen for study of those problems. The social and behavioral sciences are especially fruitful areas in which to carry out such studies because of the relative readiness of behavioral scientists to regard such investigations as highly relevant.

The assessment of value orientation is, however, beset with many conceptual and instrumental difficulties. The available instruments are severely limited in the scope of the values they purport to measure (if, indeed, they even measure these), while for the most part they are themselves too closely wedded to a particular value orientation, a bias in favor of studying what can be stated "objectively" and rendered quantitatively. Work in devising increasingly versatile assessment techniques and instruments—and even more so, further exploratory hypothesis-seeking studies—needs to be carried out in this important domain.*

Sociocultural Influences

Sociocultural factors constitute another range of determinants that affect the variables chosen for study in research. Sociocultural factors bring to research attention various problems that are relevant to the epoch and the locale. And they place an implicit,

* Such studies as Clark’s Vocational Interests of Nonprofessional Men and Morris’ Varieties of Human Values serve as background data and as examples of studies already made that are directed toward determining the patterns of relationship between value (and interest) orientations and research orientations.

and sometimes explicit, hierarchy of urgency and importance to the whole range of possible phenomena one might study scientifically. For instance, in psychiatric science in the nineteenth century, description and classification were considered of cardinal importance, and investigators were drawn to the usefulness of establishing a proper nosology, of identifying and of differentiating the emotional disorders. Of equally high importance was the parallel search for anatomical and histological change as the etiological basis for mental illness. Sixty years ago, as psychoanalysis developed, it became increasingly appropriate to examine early childhood experiences, traumas, and fantasies as the imputed etiological factors in personality disorder. During the past decade, it has become the fashion to experiment with the psychopharmacologic agents, not only to explore their usefulness as therapeutic tools in the alleviation of mental illness, but also to study the concomitant biochemical and psychological impact of these drugs on human functioning.

In England, during World War II and the Blitz, it became popular to study the effects of separation of children from their parents (Burlingham and Freud, Bowlby, etc.). The divergent development in modern psychiatry—psychoanalytic in America as against Pavlovian conditioned-reflex theory in Russia—is an outstanding example of the major molding of the direction of scientific work by the requirements and opportunities of epoch and locale. And in clinical medicine as a whole, the attention of research has been most forcefully directed toward the better detection and treatment of the disease process rather than toward more basic investigation (with less immediate therapeutic application). Today, with the anxieties of the nuclear age and its potential for our total annihilation, a distinguished panel convoked by the American Association for the Advancement of Science has urgently called for the interdisciplinary pooling of talents in order to bring a Science of Survival into being. These are only a few examples of the interplay of culture, place, and time on the problem an investigator chooses to study.
The State of Development of the Science

Research choices are also shaped by the state of development and sophistication of a scientific discipline.† This is not just a question of the overall age of the science. Nuclear physics is likewise a young science; biochemical genetics is barely a decade old. It is rather an issue of the availability of techniques and concepts adequate to the problems selected for study. It is a question of the fit between the methods (technical and conceptual) at any given developmental stage and the problems that these methods allow one to see as relevant and feasible areas for research study. To consider a just-mentioned sequence from this point of view: in the nineteenth century, anatomy, pathology, and the description and classification of syndromes were favored arenas for investigation in what we now call the behavioral sciences. Concepts and tools hardly existed for research activity outside these purviews and efforts to develop them were regarded as mystical and unscientific. The idiosyncrasies of the individual and the vagaries of his unconscious mental life became an acknowledged research area only after Freud's fundamental observations led him to new concepts and techniques that opened a new method of exploration and thus made available a new body of data.† Today, neurochemical and pharmacological research is feasible and popular but it would not be so if appropriate techniques and a widening armamentarium of pharmacological agents were not being concomitantly developed. Additionally, new techniques of teaching in disciplines crucial to research progress—for example, the newer methods of teaching mathematical theory—will help open new, different vistas in medical and psychiatric research.

This factor, the bearing of the state of conceptual and techno-

† See James B. Conant's book On Understanding Science for the spelling out of the kind of historical case study method for teaching science that is inherent in this view of the developmental process in scientific research.

† See "The Influence of Freud on American Psychology," by David Shakow and David Rapaport, for a historical discussion of how and to what extent the psychoanalytic discoveries of Freud have penetrated and gradually altered the scientific outlook and scientific work of American psychology.

logical readiness of a science upon the research choices possible in it, is also related to the differing research orientations readily discerned in differing investigative styles. An investigator may reason that when technical advances have made specific apparatuses and procedures available, he can pursue more adequately particular hypotheses about the phenomena of his field. ("Now that we have this method (tool), it is possible to work on this problem.") Such endeavor can be characterized as hypothesis-directed research. Another investigator may reason that when these (hitherto unknown) apparatuses and procedures are available, he can pursue a whole range of phenomena now made discernible for the first time. ("Since we have this method (tool), let us find out all the things we can learn from using it.") This can be characterized as instrument-directed research. Obviously the dichotomy between two such "ideal types" is in research practice neither exclusive nor sharp. Ideas can create instruments and, interactively, instruments may lead to new ideas. And although there is always the danger of a sterile search for problems to study just because they fit a newly developed instrument, the corresponding benefit is that new instruments make possible the development of new research areas.

Despite the confining tendency in the scientific community to regard as "scientific" that which it is possible to carry out at the particular point in time and place (and which therefore is in vogue among scientists) there are, of course, the exceptions to this generalization—men of genius like Freud, Darwin, Einstein, who create the innovations that are the revolutions in science.

Scientific Models and Scientific Identifications

There is yet another large category of value determinants that influence the choice of problems studied and of methods used by the scientific investigator. These are the scientific models and identifications espoused, often with cultist intensity, by their scientific adherents.

The "Exact" versus the "Inexact" Sciences

A value judgment tends to be applied to the doing of research
in the so-called "exact" sciences as compared to the "inexact" sciences, with the higher status traditionally bestowed on workers in the "exact" sciences (see in this connection the paper by Searsman and Marks in American Scientist, George Gaylord Simpson in Science, and Jerome Richfield's paper on psychoanalysis as science). Specifically, of course, research in physics (almost the prototypical science) and in chemistry has tended to carry greater distinction as scientific endeavor than research in the biological sciences or, even more, the social sciences.

This honored position of the physical sciences stems justifiably from their greater explanatory power, but this is not necessarily a direct correlate of their "exactness." Certainly, the properties of a piece of lead can be measured and dealt with exactly. But the properties of a subatomic particle can only be dealt with probabilistically. With the advent of quantum mechanics, physics has become simultaneously much less and much more exact than it was fifty years ago (depending on the area). Pari passu, in biological science, although individual growth, development, and death can only be predicted (i.e., "handled") probabilistically, population curves and mortality statistics can be graphed quite exactly. Analogously, the fate of individual electrons and individual people is quite uncertain ("inexact"), but when they are aggregated into the millions (for electrons, even higher orders of magnitude) much more "exact" prediction becomes possible.

Clearly, one need not accept the value judgment that greater respectability necessarily accrues to research activity that permits greater rigor and precision of measurement. One can set the concept of explanatory power rather than that of exactness in assessing the maturity of a science. In this sense, when psychiatry has developed further, it may (or may not) come to have explanatory power comparable to that of physics—but this will not necessarily be a function only of exactness. Making exactness an overriding consideration in scientific endeavor, rather than predictability or replicability, can lead to pseudo exactness or triviality in measurement at the expense of originality and creativity. Of more fundamental importance is the appropriateness of assessment procedures to the phenomena of interest to each scientific discipline, as part of which "exactness" may or may not appropriately develop.

"Pure" versus "Applied" Science

The medical sciences are largely applied sciences, geared to the healing mission. Psychiatric research—in contrast to research in some of the other behavioral sciences such as psychology and anthropology—has largely been medically oriented to diagnostic, prognostic, and therapeutic goals. Traditionally, psychiatric research has built on the therapeutic interview as one of its chief sources of data and of problems. Much of even the basic research in psychiatry has had its origins in the applied setting, and certain kinds of data basic in psychiatry came only out of the applied setting.

Applied research accords top priority to answering questions that have more or less identifiable and visible applications. Basic research, on the other hand, is more often pursued for its own sake, for the discovery of the nature of phenomena and of their interrelationships, without direct concern for long- or short-range applicability. Different researchers value differently the two aims, which can be labeled in a crude way as science for usefulness as against science for truth, and will hence gravitate toward different research choices.

"Medical" versus "Psychiatric" Research

The same rules of scientific procedure encompass all varieties of medical research, whether psychiatric or nonpsychiatric. They are all aspects of the medical healing art, deeply rooted in the biological tradition. There are, however, important respects in which psychiatric research differs from nonpsychiatric medical research. Though these differences may be mainly of degree, a matter of relative emphases (and, according to some, more apparent than real), it may be heuristically useful to highlight them as follows:

(a) The variables that influence human physiological functioning
are more specifiable, isolatable, measurable, and their relevance to the functioning of the physiological target systems is better worked out than their counterparts that affect human personality functioning. In the psychological realm these problems of interrelationship and relevance are more complex and are at the same time less well understood. Additionally, the sheer numbers of relevant variables and their operation at varying levels of organization (physiological, psychological, and social) further compound the problem. In psychiatric research one must be concerned not only with the impact of most of the relevant biochemical and physiological variables, but also with the effect on the individual of his own mental processes, the mental processes of other persons, of social systems and cultures, and so on.

(b) There are differences in degree of specification and of testability by empirical means of the constructs used in medical and psychiatric research. Many of the apparently most cogent and useful constructs in psychiatry are abstractions relatively far removed from directly observable and operationally testable mental events. Commonplace examples of such constructs are: ego strength, instinctual drive, frustration tolerance, superego, and psychological mindedness. The high level of complexity and the relative lack of specifiability of the empirical referents of these psychologic constructs tend to set psychiatric research apart from nonpsychiatric medical research. Such a clear distinction is of course not fully warranted, for, again, the difference is more one of degree than of kind, and the nonpsychiatric medical sciences have their share of equally complex constructs, such as cardiac reserve, immunity, and allergic diathesis.

Linked to this issue of the relative distance of constructs from observables, and to the concomitant greater potential variability and complexity of the explanations of the mental operations, is the fact that the tremendous capacities of the central nervous system to receive, code, and store information, and to produce new combinations of what has been put into it, has thus far precluded adequate description or representation by a simple digital or even an analog mode.

(c) The highly confidential nature of a patient's communications to his psychiatrist frequently precludes the full publication of research findings. Although nonpsychiatric medical data are also confidential, anonymity is usually far easier to preserve and the potential embarrassment of possible exposure far less (with some few exceptions, as in the occurrence of syphilis); consequently fewer restrictions are usually placed on the dissemination and publication of such medical case history information. The nature of psychiatric information is such that it is much more readily identifiable by the individual, by those who know him, and—perhaps in some well-known or notorious instances—by the community at large. Furthermore, one of the principal therapeutic and investigative tools in psychiatry, namely psychotherapy, traditionally does not permit the presence of a third person as a disinterested observer because of the necessarily highly confidential character of the relationship. This is a value orientation that should be reexamined and perhaps should be modified under certain circumstances; nevertheless, it prevails and therefore has an influence on what processes can be readily studied and reported in psychiatry. This situation prevails to a far lesser degree in nonpsychiatric medical research.

(d) Psychiatric research requires the concomitant consideration of social and cultural factors more than nonpsychiatric medical research. That is, the data, the concepts, and the tools of

---

§ Freud, in his "Prefatory Remarks" on the Dora Case, discusses at length (pp. 8-9) the problems involved in the publication of case history material in psychiatry in a way that both advances science and at the same time safeguards the privacy of the patient. Walsh discusses, from the opposite side, those instances and circumstances that might warrant a partial infringement of this right to confidentiality.

"Those who have worked with recording or filming therapy sessions report uniformly that therapists characteristically find such intrusion more objectionable than do patients. For discussion of the problems involved from the point of view both of implication on the therapeutic process and of invasion of privacy, see Sternberg, Chapman, and Shakow."
the social sciences perform exert considerable influence upon psychiatric research. This consideration of the interrelationships with social and cultural factors likewise occurs in areas of nonpsychiatric medical research, as, for example, in public health and epidemiology (see Halliday, *Psycho-social Medicine*), but it is less prevalent, and less necessary, than in specifically psychiatric research.

(c) An enormous quantitative difference stems from the degree to which psychiatric research depends upon the individual patient's reports of his subjective states and the psychiatrist's empathic understanding of these, as another experiencing human being. Linked to this is the fact that reliable methods have not yet been worked out for properly controlling this kind of observation.

These various differences in emphasis and in degree between psychiatric and nonpsychiatric medical research are in part perhaps reflect differences in the phenomena under scrutiny, and in part perhaps reflect differences in the traditions of the disciplines. These differences may in some areas diminish with time. Certainly psychiatric research is clearly distinguishable at this point from nonpsychiatric medical research by its greater concern with the question of individual difference and with the subjective, the personal, and the private in human functioning.

The Categories of Scientific Work

Claude Bernard's spoke in terms of three stages through which the scientific work of any discipline progresses. He designated these as observation, experimental reason, and experimentation, and he considered them the natural hierarchical evolution of a developing science. In his day much of biology, and almost all of psychiatry as well, was clearly in the first of these stages (of which the classical work of a naturalist like Louis Agassiz was a paradigm), with vigorous efforts to establish beachheads into the second and third stages (including the work of Claude Bernard himself).

Recently Helen Sargent, in discussing the hierarchy of re-

search methods in clinical psychiatric research, indicated four levels, which she termed (1) clinical (that of hypothesis finding and isolating of variables); (2) process or covariant (that of hypothesis finding and formulating); (3) factorial (that of hypothesis refinement and prediction testing); and (4) experimental (that of verification, and subsequent extension of hypotheses). Sargent explicitly disclaimed any value hierarchy in these designations: "There is no inherent assumption that exploration into frontiers is less or more valuable than road building in known territory." The disclaimer would perhaps be even more convincing if the distinction were not in terms of levels, but rather of categories that are differently appropriate to different research problems. Which category (each with its discrete methods) is appropriate will vary from problem to problem within a given field, and in terms of overall emphasis will vary from field to field. In turn, the nature of the problems in the field will determine the categories of potentially fruitful research approach.

The Philosophies of Research

The motivation for studying change and the conditions that lead to change in the phenomena of clinical psychiatry might be seen, most generally, as rooted in the contribution that the research undertaking promises toward the extension and the enhanced precision of man's knowledge of himself. Such a global view was more congenial (and more satisfying) to nineteenth-century thought concerning the role and goal of science than it is to more cautious present-day assumptions about the inevitable limitations of science and scientific method. Within this overall framework, however, there are more specific values and rewards, which relate directly to the goals of scientific endeavor in itself, quite apart from the more personal (or less "noble") motives for their existence.

These specific values inherent in the motivation for scientific research may be distinguished as those of (1) elegance, (2) confidence, and (3) utility. The corresponding frames of reference are the concerns with (1) meaning, (2) instrumental precision,
and (3) applicability. To specify more exactly the links between these values and the preoccupations and procedures of science, the first—elegance—is clearly a central element in the worth assigned to a body of theory or a theoretical system. High explanatory value and parsimony are hallmarks of a successful (i.e., elegant) formal statement. The physical sciences have been envied, but for good reasons not matched by the biological sciences, in this quality of elegance of their most powerful theoretical tools. Confidence, while clearly linked with the ability to predict successfully, is also associated in practice with the ability to manipulate precisely the situation under study, whether this manipulation be by statistical techniques, by refined instrumentation, or by rigorous control of the conditions of the research situation itself. It is centrally related to, and valued because of, the reliability that can be claimed for the inferences derived from the research. Finally, as regards utility, while it might be viewed as suspect or limited as a sole goal in its own right, it would certainly not be disavowed as a legitimate and important by-product of scientific effort. Its specific claim to consideration is its high operational value, which is in turn linked with the values and goals of the social system. If adherence to a particular body of theory has acquired value within the social system, a conflict may develop between the utility for the social system itself of research findings supporting a valued theoretical position and the more pragmatic and more directly applicable utility of other findings that may not be congruent with this theoretical position. As a result, utility is paradoxically the most complex and difficult to categorize of the three sets of values.

Ideally every research undertaking would fully realize each of these three kinds of values in its results. This is, however, rarely the case in practice, and the investigator, in selecting a research issue in which he proposes to make a large personal investment,

---

To give a commonplace illustration, research with the psychopharmacologic agents in relation to behavioral change (and improvement) in hospitalized psychiatric patients rates very high in utility, only medium in confidence, and low indeed in elegance.
Determinants of Research Choices

Depending on the particular values accorded predominant weight, particular part-aspects of the subject and of the environment will be selected for formal scrutiny in the formalization and execution of a specific piece of research. While these part-aspects will overlap with the part-aspects that would be selected on the basis of an alternative set of value priorities, they are unlikely to be completely coextensive. This is in itself a fertile source for the differences in the conclusions that may emerge when the value orientations underlying the respective investigations differ.

The formulation of a research hypothesis or the explicit statement of a research problem includes among its functions the protection of the researcher against the overwhelming complexity presented by the living, interacting organism. How much protection is optimal cannot be summarily established or stated. Among the considerations involved are the following: (1) the degree of existing warrant for belief in the antecedent assumptions that have led to the particular narrowing of the field of attention and inquiry adopted by the research design selected; (2) the limits of flexibility of the techniques, instruments, and data-handling resources available to the research undertaking; (3) the degree of openness and uncertainty that the researcher himself can both tolerate and utilize effectively; (4) the type of “answer” that the research is intended to provide.

Only the first of these four considerations addresses itself specifically to the question of what would be optimal (in the sense of most congruent) with the research question asked and the state of existing knowledge in the field. The three additional considerations are concerned with types of limitations either inherent in the research situation itself or consequent upon the arbitrary restrictions placed on inquiry by choices that—in following certain directions—preclude the pursuit of other choices. It is in these three factors that the value orientations of the investigator will be most directly reflected, and in which these value orientations will play a determining role in limiting the alternatives acceptable to the investigator and the priorities that he assigns to his choices.

It is at this point that the “competing” value orientations of elegance, confidence, and utility operate. Each of them operates on at least three distinguishable levels of inquiry and scrutiny.

The first level is that of priorities in the selection of problem and methodology. It is concerned with the general characteristics of those aspects of the total situation that tend to be selected in formulating the research question and in selecting methods for exploring this question. Within the value orientation of elegance, choices of problem and method are made on the first level in terms of their congruence with a conceptual framework, a body of theory. Within the value orientation of confidence, choices are made on this level in terms of their amenability to precise manipulation. And within the value orientation of utility, these choices are made on the basis of their congruence with social system values.

The second level of inquiry and scrutiny is that of the justification of the research and its results. It indicates the formal grounds on which the investigator rests his claim for the particular value of his research and his research findings. Within the value orientation of elegance he is concerned with the high putative explanatory value of his work and his findings. Within the value orientation of confidence he is primarily concerned with their high nominal reliability. Within the value orientation of utility his chief concern is with their high nominal operational or utilitarian value. Any judgment by an investigator made from the point of view of this level of inquiry on the worth of research conducted with a different basic value orientation is likely to question the degree to which the “really” important issues are considered and adequately dealt with. For example, the individual who predominantly values elegance will tend to question the pertinence of “excessive” stress on the reliability of the findings or on their high operational value; similarly for individuals looking from each of the vantage points at the others.

The third level is that of the basic value orientation in the three reference frames, the values of elegance, confidence, and utility. From this level, judgments made on research stemming
from each of the differing value orientations will almost certainly be seen as seriously lacking in both salience and relevance. As one ascends the levels, the orientations diverge. It is more possible to ask from the vantage point of level one of each reference frame (the level of priorities in the selection of problem and methodology) whether those aspects selected from the total situation and the proposed research techniques give sufficient minimal consideration to those elements at the same level in the other two reference frames, which would unquestionably be incorporated in any competent research undertaken with either of the alternative basic value orientations as its point of departure.

In making judgments of salience or relevancy, both the reference frame from which the research was undertaken and the reference frame from which the judgment is made need to be specified if confusion is to be avoided, since each predilective position molds the bases of judgments made in reference to procedures or conclusions originating from another predilective position. To understand fully the relatednesses and interdependencies involved requires in each instance explicit consideration of two questions: First, from what vantage point can judgments of relevancy and saliency of specific research involving any of the value orientations be undertaken with greatest profit? Second, what are the most probable sources of strength, on the one hand, and of limitation and distortion, on the other, of research conducted under the predominant influence of each of the three value orientations?

The position any investigator occupies at the third, the most abstract, level of inquiry, the level designating the basic value orientation itself (elegance, confidence, utility) is seldom entirely a matter of conscious decision. Level two, the level of justification of the research and its results, represents a set of attitudes that are nominally determined by formal reason and the exercise of judgment and hence are presumably subject to revision on the basis of convincing demonstration and broad experience. The “nonrational” contribution to this level from the third level runs the risk of being denied or slighted. When this occurs, the illusion of objectivity and impartial judgment—supported by suitable arguments that are unassailable by counterpropositions based on alternative premises—is likely to limit severely the possibility that insights and techniques deriving from other value orientations may become available for application to and critical assessment of the given research position.

This is not to suggest that investigators as a class can be neatly divided into three groups that coincide with these three basic value orientations. The individual who can be completely categorized as a representative of only one of these three value orientations is fortunately the rare exception. While such single-mindedness would not preclude valuable research contributions by the highly gifted individual, it would be at the price of some constriction of those endowments and would impose some limitation on the individual’s capacity to place the implications of his own work into the most meaningful perspective.

Each reference frame has, then, in addition to its inherent strength, its characteristic limitations and vulnerabilities. The identification of these could, ideally at least, serve to forewarn the investigator against the pitfalls that are most likely to beset his particular path. It would be naive to assume that this ideal can be readily realized, since the nonobjective component inherent in the predilective position adopted by each investigator provides a barrier that is not readily penetrated either by the investigator himself or by the examination and discussion of the formal elements of the situation. In addition, the more completely the predilective position of the investigator is confined to a single value orientation, the less the likelihood that the investigator can afford to recognize either alternative positions or the degree to which his own judgments are unilaterally determined.

An example given by Peter Kapitza will serve as final illustration and as summary of the many considerations, extrascientific as well as scientific, that lead to research choices. The law of the indestructibility of matter has been one of the fundaments of classical physics. A chemist burned carbon and carefully carried his calculations to the tenth decimal place. He expectedly
confirmed the law that the product of the combustibles equals the product of the initials. Had he gone to the eleventh place he might have discovered a minute discrepancy and thus a clue to the law of the transformation of matter into energy. We can only speculate as to what kinds of considerations determined his choice of the point at which to stop in his calculations. Is it possible that he went two decimals beyond the tenth and then dropped the last two measurements, which didn’t fit the existing theory, as reflecting instrumental error?

2

CONCEPTS OF CHANGE

The concept of change implies a difference between two states. Change has been defined as “a concomitant variation in time and some other respect; a departure from a norm; a deviation from an established character, sequence or condition; a divergence from uniformity or constancy in any quality, quantity or degree.”† All these kinds of change have relevance for one or another of the kinds of problems placed under scientific scrutiny in psychiatry. The following considerations have to do with some of the important assumptions and logical bases on which psychiatric research necessarily rests.

Context of Change

Event and Its Context. Any assessment of change occurs within a context in relation to which the change may be clearly delineated or may be obscured, exaggerated, or otherwise distorted. In psychiatric research, there is particular danger of distortion. For example, the specific culture of a chronic ward in a large mental hospital might camouflage a change in the intensity of psychotic illness through the characteristic operations of the ward culture’s value system. Thus, a patient might continue to behave in a withdrawn fashion, despite an increased capacity for more social behavior, were he in another kind of social field, one that did not condone withdrawal. In psychiatric research, the need to define the relevant context in which a given research

† WEBSTER’S SECOND NEW INTERNATIONAL DICTIONARY, G. & C. MERVISS CO., SPRINGFIELD, MASS.
problem is being studied takes on special significance. In other research areas it is more often possible to control for context by standardization; but in psychiatric research, we cannot standardize or control an interpersonal field in the way or to the extent that temperature and pressure can be controlled. When we compare psychiatric treatments, the problem of context may become particularly acute. Many forces within the social field may produce effects indistinguishable from those under specific study. For example, the introduction of a psychoactive drug into a treatment program in an institutional setting may concomitantly generate social forces within the treatment setting that move patients in the same (or in the opposite) direction as the pharmacological action of the drug, or in the same direction with some patients and in the opposite direction with others. And even the pharmacological action of the drug itself can be altered by the set. The need for an adequate statement of the relevant forces operating in the research context, as well as for the delineation of the specific object of research study, is clear. In fact, the widely discrepant drug responses reported in the psychiatric literature may reflect in part an inadequate delineation of the contexts in which experiments have been carried out.†

This important distinction between object and field in psychiatric research may be difficult to make precisely, since the definition of the object itself is often elusive. Thus, the categories of the classical diagnostic groups in psychiatry are definable only with reference to modal populations. But when a study is carried out on small groups of patients, not necessarily the most representative of a given diagnostic category, the differences between individuals in the group are often of greater import than the mean differences from one diagnostic category to another. Quite frequently one notices great concern over the delineation of the precise characteristics of the object under study, with relative neglect of the necessary complementary interest in the characteristics of the context in which the study is being carried out. We have detailed criteria for matching patients on a variety of demographic and individual dimensions, but we have paid little attention to establishing criteria for comparing or matching the contexts in which patient populations are placed. Delineation of context has most relevance for research on the relationship between clinical phenomena and other variables; it has less for other areas of psychological research, where standardization of the research context appears more feasible—and is hence more practiced.

Baseline. By definition, change involves a difference between two states in time and some other concomitant variant(s). The implication of this definition is that the first state is specifiable and that change from this initial state can be seen along some standard rule or dimension. A difficult problem in psychiatric research, however, is the instability of the baseline or initial state. Thus, an assessment of change in the psychopathology of an adolescent must consider the developmental dimension; any measure of intensified or attenuated psychopathology must be made in terms of a deviation from a moving base point. Stated alternatively, it is not always easy to tell whether a young patient has become sicker or just more adolescent. There is also confusion as to whether these two kinds of change should be seen as identical, similar, or different. The same problem occurs in studying other populations as well—for example, aging patients.

Because evolving psychopathology may be confounded with natural growth and development or with aging, it is difficult to establish a suitable standard for measuring change. Furthermore, since it is not clear what any given person’s “natural” rate of growth and development is, it is not possible to distinguish psychopathological processes from growth processes by their differing rates of change alone. The process of growth through the adolescent period (or, for that matter, through any life period) in all likelihood has no fixed and even rate of change. It seems to occur in as yet unpredictable spurts, intermingled with

†This discrepancy of response is smaller with the phenothiazines, greater with the antidepressants. See Cole et al. for good reviews of the psychopharmacological literature on drug treatment and for an idea of the range of results often obtained with these constant chemical agents.
periods of relative quiescence. Whether a specific treatment intervention initiates a period of growth or whether growth might have occurred in any event presents a difficult problem for research control and for the setting of adequate standards for assessing changes in psychopathological states. In deciding whether an adolescent has improved over the course of two years of treatment, it is not ordinarily sufficient merely to compare his behavior at 14 with his behavior at 16. Rather, one must compare the appropriateness of the behavior at 14, in terms of reasonable standards for a 14-year-old, with the appropriateness of the behavior at 16, in terms of reasonable standards for a 16-year-old. What is behaviorally appropriate for the 14-year-old is no longer appropriate for the 16-year-old.

Thus, in evaluating change the psychiatric researcher must be prepared for the subject to become different in many ways other than the particular aspect under consideration, and for these other ways to confuse the dimension under study. One must not hope for a simple, single standard by which to measure change, since behavioral standards change with age and time. This is not to say that change cannot be systematically studied, but our research must consider this complex interplay of growth and illness processes. The basic question is, how much of the classical empirical systems can be appropriately retained and what modifications need be devised to enable systematic study of the problems of clinical psychiatric research?

Kinds of Change

Alteration versus Substitution. An alteration is a change in the quantity or quality of a thing; a substitution is the bringing in of a new thing in place of the old. In evaluating treatment outcome in psychiatry, we sometimes have difficulty in distinguishing between alteration and substitution. The shift in a particular defensive constellation by which a patient deals with anxiety occurs often in the form of the emergence of some other, differently constituted, defensive constellation. It is relatively easy to say that a patient is less paranoid, and therefore less sick, where this disposition is the central relevant aspect of his illness. It is more difficult, when the configuration of the patient's defenses changes in kind as well as intensity, to say that a patient who was paranoid but is now depressed is healthier now than before. For those who feel that the mechanism of blame connects paranoia and depression, this change is an alteration. For those who don't subscribe to this connection, the change is a substitution. The fact that there can be two such viewpoints indicates the difficulty in differentiating one kind of change from the other. Faced with this dilemma, an investigator finds himself reaching for a health-sickness dimension that is central to a given patient's illness and yet has remained constant through the course of change. Such a stable dimension may be difficult to delineate, and yet doing so is fundamental to being able to assess the changing clinical state on an overall health-sickness dimension.

Qualitative versus Quantitative. Although quantification is generally accepted as one of the most powerful operations of science, and as a fundamental condition for the rigorous testing of hypotheses regarding the relationships between variables, many of the phenomena most relevant to psychiatry do not occur along readily quantifiable dimensions. These are phenomena that, if reduced to "more or less" terms, lose something salient.

Because of the difficulties in this area there has been a tendency in some quarters to deny both the possibility and the importance of quantification. At the opposite extreme, some workers choose to deal only with those phenomena that can be rigorously quantified even when their relevance is tangential or trivial. In recent years, the quantification of complex behavioral variables such as affect intensity has been approached and it may be hoped that such efforts will be expanded by physiological, linguistic, and psychological techniques. In theory at least, any observable event may be quantified, and inferences derived from combinations of such observables may also be quantified. But the kinds of changes that often occur in psychiatric phenomena make the technology of quantification complex and difficult.

Change in Time. Much psychiatric research involves observing
change over time. The frequency of observations and the length of time over which one observes are often critical in determining what one sees. Too frequent observations may result in redundancy and a falsely increased \( n \) in calculating probability. Too few observations may result in error due to inadequate assessment of the variance from sources other than the experimental source. For example, in a study of the effect of treatment interventions upon a schizophrenic patient the results may be different if the total time of the study were six weeks rather than six months. Again, if observations are made on the patient every 30 minutes rather than less often, there would be a great number of observations, but many would be the same, that is, redundant. Also, if observations are made only at the beginning and at the end of the study period, one might not be able to differentiate between a "natural" fluctuation, a stable change in the patient, or a change in the frequency of fluctuations. That is, the time dimension of observation needs to be adjusted to the kind of change one is trying to observe.

Change in Frequency. Phenomena of interest can also change in frequency, that is, in the number of occurrences per unit of time. Study of changes in frequency introduces problems in proper sampling. The time intervals studied must be of sufficient length to permit accurate measurement of frequency and must be spaced so that changes in frequency can be validly ascertained. There are also problems of weighting and of meaning. The seriousness, the magnitude, or the saliency of a phenomenon is not necessarily encompassed in an altered frequency count. A single, explosive instance may outweigh in importance and meaning many repetitive mild instances. Likewise, the degree of improvement in the psychological state of the handwashing obsessional is not to be equated just with the linear reduction in handwashing frequency.

Change Through Growth and Development. Change along the growth and development continuum presents a special problem already discussed, namely, that of unstable rate of change, as well as the questionable linearity in the relationship between amount of change and time. For example, the differences between an infant 3 months old and another 6 months old would be much greater than the differences between an adult 20 years old and another adult 20 years and 3 months old along most dimensions. The amount of developmental change would ordinarily show much greater quantity per unit of time in the earlier years of life than in the more stable years. Such problems are compounded further by the similarity often present between growth processes and healing processes.

Observables versus Infullerables. Efforts to combine observations of behavior into unified meaning structures for purposes of clearer communication and understanding have resulted in the use of many constructs that refer not to observables but rather to inferables. These are states inferred by the presence of combinations of observed phenomena that may or may not have an invariant relationship to the inferred state. Thus, assessing the presence and the amount of anxiety as an effect in a patient is a complex judgmental act made by inference from varying combinations of manifest cues. Such a statement of anxiety may then be said to pertain to an inferable, based on the assessment of observables, as organized within a theoretical frame of reference that gives them meaning. Change can be separately assessed in the observable (tremor) and in the linked inferable (anxiety).

Occurrents versus Dispositionals. A further distinction is that between occurrent terms and dispositional terms—for example, the distinction between the roundness of a plate, which is a constant existing property (occurrent), and its fragility, which is its proneness to break when subjected to certain operations such as dropping (dispositional). In the study of personality disorder and its modification by psychotherapy, the field of observation is only a fraction of the total interpersonal field of the patient, yet generalizations are made from the time-limited dyadic office relationship concerning capacities and dispositions both within and beyond the treatment situation itself. These extrapolations often make use of dispositional terms, as, for example, anxiety tolerance, which is a dispositional statement of the intensity of the
anxiety that would become manifest under particular conditions of drive pressure and ego control. Other terms, like anxious, can at times be occurrent (as in the statement that someone is anxious, meaning that he is at this moment manifesting various recognizable signs of anxiety) and at times dispositional (as in the statement that someone is anxious, meaning that he is chronically beset by a readiness to anxiety, even though at the moment he is calmly asleep). Ordinarily, occurrent properties are more easily available for assessment than dispositional properties. In clinical research, models must be devised that will provide a basis for valid statements of status and of changed status in regard to both kinds of terms. To assess anxiety as an occurrent state will require devices different from those required to assess anxiety as a dispositional state.

Dimensions of Change

We will consider a variety of dimensions of change in order to introduce some of the assessment problems attending each dimension. These dimensions are not mutually exclusive sets, nor is every one of them applicable or significant in each instance of assessed change. Depending on the instance, the consequences of ignoring one or more of these dimensions can be quite different.

Occurrence. The primary dimension of change is the judgment of the occurrence of change, of presence or absence, yes-no, on-off, and so on. Every statement of direction, amount, frequency, and the like contains an implicit assertion that change has occurred.

Direction. Direction adds the next simplest dimension to change, with no necessary implications for intensity, rate, duration, or sequence. It does involve coordinates. Direction can be toward mature or immature behavior, toward progressive or regressive adaptation, toward growth and differentiation or the lack of these, and so on. For specified purposes these directions are often termed better or worse. In each instance, direction can be discerned by the relationship in time of the events to each other. Thus, sequential readings of the clinical state of a patient can yield directional information with regard to whether the patient is becoming healthier or sicker. A stricter, more specific quantification of amount of movement in the given direction often may not be possible or even useful. It is implicit in a statement of directional change that there is some referent—either the same event looked at on subsequent occasions or a comparable event looked at simultaneously. Thus, the measurement of amount can be based on a standard derived from prior observations of the same object of study (an ipsitive standard), or it can be based on a norm derived from observations of similar events in a similar situation (a normative standard).

Magnitude. In psychiatric research, what constitutes magnitude or amount is not necessarily self-evident. It may refer to the number of times an event occurs, or to the intensity with which it occurs, or to a combination of both. A statement of the magnitude of change implies that change has occurred and in a given direction. It adds the additional dimension to direction (i.e., healthier-sicker) of amount (i.e., how much healthier, how much sicker). It likewise implies a standard of reference, again either ipsitive or normative.

Rate. Rate of change is the amount of change per unit of time. Growth, healing, and other processes considered in psychiatric practice are frequently uneven in rate, and measures of rate have to take this unevenness into account. A basic problem is distinguishing between the rate of change in the total field and the rate of change in the variable under consideration within the field. Frequently, both are changing. Thus, if one is interested in the rate of alteration in the psychopathological state of an adolescent, one has the problem of assessing the rate of change in the maturational process and of comparing this to the rate of change seen in the pathological process. One has then to decide whether the two should be simply "subtracted," one from the other; or whether, by parcelling out the growth rate from the rate of change in the pathological state, one is not destroying an essential aspect of the evolution in the psychopathological state.
Duration. The issue of how durable a change can be or may be is often a thorny one in psychiatry. Successive examinations of the selected end state are crucial to this issue. Conclusions about the permanence of change depend in part upon the length of observational time span one chooses. Many of the most salient distinctions in regard to the theoretical basis for observed changes through the course of psychotherapy derive in large measure from considerations of relative permanence or duration (e.g., the distinction between changes based on conflict resolution and structural alterations in the ego and changes based on transference cure.) An example of this problem of duration, which is very crucial to research strategy, is that of the appropriate timing of follow-up study after psychiatric treatment. The time chosen for follow-up study frequently has major effects on what one observes and on the conclusions that are drawn from such observations.

Reversibility. Reversibility of change is a factor closely linked to that of duration. Irreversible changes are, by definition, of permanent duration. But little in our field is in this sense truly irreversible. Rather we are dealing constantly with changes in the plasticity or lability of a system, as against its rigidity or its brittleness. Freud spoke of the situations in psychoanalysis where change that is readily induced is often equally unstable, whereas more hard-won gains often have correspondingly greater durability. The same system usually varies over time in its susceptibility to change. At times of crisis or stress, an individual will tend to be more susceptible to change in both disruptive and reintegrative directions; that is, change will be more reversible. On the other hand, symptoms and states may differ inherently both in inertia and in momentum.

Discernibility. Changes in discernibility can often be confused with changes in degree, magnitude, or even occurrence of a phenomenon. Thus, an enhanced capacity for circumscribing his psychopathology and for simulating may lead to a decreased visibility in the paranoia of a patient, with no significant alteration in its intensity or pervasiveness. Yet, based on the diminished discernibility, an erroneous judgment of significant improvement may be made.

Sequence. This dimension of change has relevance in psychiatric problems in a variety of ways. If one can establish a regularly recurrent sequence or pattern of change, additional information is often revealed about the nature and meaning of the phenomenon under study. For example, in a study of the social contact patterns of patients going through acute psychotic episodes, the number of people the patient socialized with per social occasion and the relationship of this pattern to the psychotic process over time were recorded. Some patients withdrew progressively from social interactions by first decreasing the number of people per social occasion from more than one other person (a group) to just one person (a pair), and from this paired kind of social occasion withdrew to complete isolation. This isolation continued until the start of resocialization, at which time there was a reversal of the process, so that the patient became involved in paired social contacts at the first stage of resocializing and subsequently in group social contacts. This sequence of patterns of social interactions highlights an interpersonal dimension of the process of psychotic withdrawal in certain patients. On logical grounds alone a sequence involving, first, the presumed intimacy of pairs, next, the diffusion of attachments over a group, and then isolation, with the reverse sequence in the recovery phase, could be just as likely. This would lead to different conclusions concerning the nature of the stresses in the social field as these relate to psychotic withdrawal. In other words, establishing the sequence of events is an important way of determining major aspects of the process being observed. The establishing of a sequence depends in large part on the capacity to observe an object over sufficient time with one or more of its aspects changing while other aspects of the object and of the research field remain stable.

Locus. We have already mentioned a variety of difficulties in defining the object under study, the context in which that object is evolving or changing, and the standard against which that change
may be measured or observed. The difficulties in defining these three areas often lead to a consequent difficulty in identifying the proper locus of change. The participant observer in a research field must not assume that the identification of a change necessarily locates that change in the object being observed. The change may appear to be in the object as a result of an illusion, when, in fact, the change is really either in the observer or in the field in which the object exists. This problem is compounded by the likelihood of movement in the field as a characteristic of the field, and of movement in the observer. The latter may occur in terms of the observer’s increasing experience and sensitivity over time and/or his increasing cognitive and affective involvement with the object or the field under study. This confluence of growth in the object, of alterations in the social forces in the matrix within which the object rests, and of shifts over time in the observer presents difficulties that must be recognized and taken into account in the assessment of change. There is also the possibility of change in the relationship between object and observer, and of change in the instrument with which the observer makes the observation. Each of these can be the apparent locus of an observed change, seemingly clearly manifest “out there.”

We have summarized some of the issues that derive from considerations of the assumptions implicit in the concept of change and of its assessment in psychiatric research. This statement of issues should not be misconstrued. A complex machinery of methodological analysis is not necessarily vital for every seemingly simple research question. Rather, these considerations should alert one to the range of difficulties that may be attendant upon any method of assessing change in psychiatric phenomena. Such difficulties should not forestall psychiatric research; they should help make it more humble on the one hand and more sophisticated on the other.

3

FACTORS INFLUENCING THE ASSESSMENT OF CHANGE

Up to this point we have considered, first, the variety of human factors and value orientations in what have conventionally been thought of as the value-free matters of choice of research problem and the choices, within the problem, of the elements to assess. Following this, the concepts of change that set the framework for our research inquiries were examined—change in relation to its context and its baseline, the various kinds of change available for study, and the dimensions along which these changes can be studied. We now turn to the various factors in the research situation itself that may influence the assessment of change. These include the factors in the research field, the reference frames and concepts that guide it, and the arrangements of time and space in which it is embedded.

The Subject

The subject often exerts a significant indirect influence on the assessment of change. Investigators in psychological science, and in physical and biological science as well, are usually alert to the well-known and well-studied pitfalls of observer bias. We may be more prone to forget the equal possibilities for distortion introduced by the awarenesses and expectations of the observed, and by the affective relationship between observed and observer, upon the observations made by the observer. At the simplest level, if the patient under study knows that you are interested, for example, in the frequency, intensity, duration, and total patterning of his headaches, and if your interest matters to him, then he may have (or more precisely may report) more or less headaches,
and in differing configurations, than he might otherwise. And in the nature of human relationships such interest nearly always does matter, in significant and idiosyncratic ways.

This phenomenon may be introduced into research assessment in our field at different points: in the initial set of the subject toward the posed task or inquiry, in the experimental instructions given the subject, in procedures designed to guard against or to exploit suggestibility, in the rapport between observer and observed, in the emergence of transference and countertransference, in the “demand-character of the experiment,” and so on. Some of these points are touched on in more detail in other parts of this report. In research, no less than in clinical practice, powerful affective currents within the observed and between him and the observer may modify decisively what is observable. These often cannot, and usually need not, be eliminated from the research situation; if they are recognized and their contribution is carefully assessed. The danger lies in the prior assumption that they will not be present.

The Observer

The possible influence of the observer himself on the assessment of change has been widely recognized in all the sciences under such rubrics as “the personal equation,” observer bias, countertransference, and the like. Of special importance to psychiatric research is the fact that the development of appropriate instruments which substitute for or provide adjuncts to the observer is often more difficult than in other sciences. At the same time the criteria for the specification of clinical entities and psychic processes are less precise, less discrete, and less agreed-upon than in many other fields. For many of the clinical phenomena with which we are concerned, our best instrument—and, in fact, the only one that at present can cope with their complexity, subtlety, and ambiguity—is the clinical understanding of the trained and experienced clinician. And even in those instances where more objective and public evaluative criteria and instruments do exist, often the clinician is still an indispensable agent for providing coherence and meaning to the the psychiatric phenomena under study.

Given these circumstances, the influence of the investigator on the field of study has a significance sometimes beyond reckoning. A variety of techniques have been devised to minimize the potential distorting component of this influence. One such technique that can be used in some research situations is the double-blind technique, in which the investigator seeks to reduce bias through maintaining ignorance of the influencing agent introduced. The use of multiple observers making independent observations, with or without later group discussion and attempts toward group consensus, can be considered for other situations in which the double-blind technique is not applicable.

There is an understandable temptation toward the use of pseudo-precise and pseudo-objective methods (e.g., the inappropriate use of rating scales) that often have achieved widespread acceptance with too little regard for the sacrifice in clinical richness or meaning that may be involved in the particular use. Frequently, the use of a rating scale has been thought to be more “objective” just because it is a rating scale, while simple overall clinical judgments (in effect, global ratings) have been rejected as overly subjective and, in consequence, hopelessly biased. Such an equation of numbers with objectivity ignores the possibility that the subjective element in the single global rating by the observer may be as strongly represented in fifty circumscribed ratings made by the same observer but clothed in the seeming objectivity of the rating scale. This statement does not, of course, imply that properly devised and applied rating scales do not have an important basic usefulness in clinical research in psychiatry.

The Concepts

It is the investigator’s task to establish the conceptual framework that will guide the research and to assess, in relation to this framework, the suitability and adequacy of the techniques employed. Clearly, the conceptual framework that is adopted—whether, for example, psychoanalytic or conditioned-reflex theo-
ry—importantly influences the problems chosen for study, the kinds of data the study will yield, and even the kinds of conclusions that can be drawn from the study and the ways in which these conclusions will fit into the already accrued body of knowledge. Research that is well thought out and designed either uses concepts and constructs that are common currency and employs them in their familiar sense, so that further specification and definition are unnecessary, or specifies and defines its concepts and constructs when they are being used in novel or in local, idiosyncratic ways. Further, the assumptions that underlie the research endeavor and the working hypotheses that guide it are spelled out in a manner that derives consistently from the concepts and constructs of the theoretical position espoused. Finally, the selection of variables chosen for study and of methods for their study follows logically from the foregoing steps. Obviously, the many kinds of "nonscientific" considerations that help determine research predictions can importantly affect each of these choice points, sometimes in ways not completely congruent with the "scientific" requirements of the problem under research study.

Time and Sequence

There is an important and obvious relationship between the kind of change to be identified and the time unit chosen in which to identify it. Some variables of interest in psychiatric research change rapidly, within minutes, and some change, if at all, gradually and almost imperceptibly. To assess the former, observations need to be made frequently and rapidly; fluctuating affective states are an example. For the latter, a long interval and a perspective on the entire panorama of a process are necessary to the assessment of the change; examples are changes in the patterning of defenses; enduring changes in structural character configurations, such as a masochistic disposition, or in complex inferential constructs, such as the tolerance for anxiety.

Difficulties can arise in psychiatric research when there is not a proper congruence among the phenomena under study, the variables chosen for the study, and the time span within which the study is to be carried out. For example, one set of criteria and one time dimension are required to assess the effectiveness of tranquilizing drugs in controlling the disturbed motor and affective behavior of an acutely ill psychotic individual. The assessment of the effect of psychoanalysis in altering the propensity for disordered motor and affective outbursts in the severe impulse-ridden character requires a different set of criteria and an entirely different time dimension.

Temporal and phasic differences in the assessability of variables are linked inevitably (though not necessarily in a unilinear way) to the complexity of the variables. This must be taken into account in selecting research strategies. Many of the most important judgments that are repeatedly attempted in psychiatric practice and research involve the evaluation of change in such complex and highly inferential constructs as neurotic character structure, ego strength, or schizophrenic alienation. Judgments of just this sort are in fact the ones most frequently made in attempting to assess changes consequent upon psychotherapy or other psychiatric treatments. Quite often it is not realized that judgments concerning such complex constructs are inferences derived from many separate observations and from the assessments of a host of behavior variables over time, and that these are all organized within a theoretical frame of reference which gives them meaning and establishes their relationships. Each of the behavioral variables contributing to the overall judgment may have a different time duration and sequence for a complete cycle, phase, or event. In other words, it must be kept clearly in mind that although each of the different characteristic aspects of schizophrenic disorganization and thought disorder may be separately assessable, assessment will have to be made on scales of magnitude and over time spans that are not necessarily equated or even closely comparable. In addition, the relevance of the separate elements to the overall integrated functioning will also differ.
Space and Field

The field of observation likewise influences judgments upon the object under study when the forces within the field counteract, modify, or exaggerate in unrecognized ways the effects produced by those factors that are being specifically studied. For example, if one is studying the effectiveness of individual psychotherapy with a particular class of patient or of illness, and the patients under study are living within the therapeutic milieu of an active treatment hospital, the effect of the social field might confound the effect of the psychotherapeutic effort. Both mobilize forces that move at times in the same direction or at times exert subtly interfering—or even openly opposing—effects upon one another. It may become very difficult to make distinctions between these forces and to assess their respective contributions to the changes observed. Yet this kind of potential impact of the field upon the assessment of change is too often slighted.

This delineation of the impact of the field upon the assessments of change in the objects of study depends in turn upon how the appropriate limits of the field are conceptualized and defined. In experimental research this is conventionally thought of as the specification of the potential universe of study and appropriate sampling from this universe. Assessment can be of uniquely defined individuals or of specified small or large groups: the psychotherapy group of two, patient and therapist; or the larger groups of family, culture, or subculture; and the restricted groups demarcated by age, or sex, or race.

In each instance, the sampling must consider the setting of the individual or the group in its situational context. For example, a regressed schizophrenic individual may be communicative and coherent when seen by his therapist in individual psychotherapy. The same individual may be uncommunicative to the point of being monosyllabic or even mute when seen by the psychotherapist in a joint interview with his family members, even at a comparable point in time. This situation, of course, could just as feasibly be reversed in the two settings. Furthermore, members of the schizophrenic patient’s family may characteristically behave quite differently with each other when he is among them than when he is not. And similarly, any individual may behave differently—in some respects, at least—when he is a member of a very small group (the dyadic psychotherapy group), or a larger group (the whole family), a disciplined group (coworkers or schoolmates), or an undisciplined group (a mob). In each instance it is necessary to select the proper field and context for the appropriate assessment of variables and of change in variables in the research endeavor.

§ See Alfred H. Sturton and Morris S. Schwartz: The Mental Hospital, for a pioneering discussion of the many possible vicissitudes of this kind of interplay.
study of the logic of measurement by remarking that "psychologists should indeed not rush into the use of physical terms if there is nothing else to be gained but a share in the prestige of the physical sciences." Aristotle similarly proposed—over two thousand years ago—that "it is the mark of an educated man to look for precision in each class of things just so far as the nature of the subject admits . . . ." Some things are but falsified if you try to be precise about them—indeed the concept of precision itself may stand elaboration; it does not necessarily mean the same thing in psychological science that it means in physical science.

Our efforts to measure change, in both psychiatric therapy and research, reveal a need to reconsider our notions of both measurement and precision. To begin with, the usual assumption that measurement necessarily presupposes the methods and concepts of physical measurement is questionable. It is at least as arbitrary as the assumption that the meaning of science can be exhaustively analyzed by a description of the procedures of the physicist. Just as decisions about methods of investigation depend upon the nature of the data sought, the success of any effort at measurement depends upon the relations between the method employed and the empirical data of the situation under study. The theory of measurement did not develop with the logical independence from factual considerations possessed by other fields within mathematics.† The properties of the real objects we experience in no way limit the development of a geometry, but the

---

* In fact, questions may even be extended well beyond this sort of consideration. A recent survey of various positions on the nature of measurement begins with the observation that "there is anything but common agreement among scientists and philosophers of science as to just what measurement is and how it should be performed." (Italics ours.)

† We mean here to distinguish the logical independence from factual considerations possessed by mathematics from the concept of historical or causal independence that is not possessed by mathematics. The kind of independence intended here is only independence of logic and not of those experiences undoubtedly involved in stages of concept formation. Any statement is logically independent of fact when no kind of experience would be relevant to its confirmation or disconfirmation. In this sense, the propositions of mathematics are regarded as propositions of logic rather than as factual statements.
conception of measurement itself does entail a relationship of congruence, fit, or "isomorphism" between the features of the measuring instrument and the properties of the phenomena to be measured.

Not all phenomena or properties, therefore, will be measurable in the same sense. It may well be that some properties may not be measurable in any sense—at least, not unless the term measurable is somehow made to include some distinctive forms of discrimination and classification not ordinarily intended in its current use. Thus, if unconscious anxiety or guilt is to be successfully measured, it seems likely that this will require several successive steps. There would need to be clarification of the differences between these psychological qualities or states and such physical qualities as weight or hardness; and clarification, as well, of the differences in what is meant when one "measures" psychological, as against physical, qualities. Further, there would need to be a specification of the nature of the differences even between two purely physical qualities, such as weight and hardness, with due recognition of the logical principles involved in the diverse methods for measuring them. And finally, when the peculiar data of psychiatry are considered in the light of the formal conditions of measurement, there would need to be a willingness on the part of the psychiatrist either to abandon some of the measurements he now hopes to make or possibly to assist in the development of the sort of mathematics that will be isomorphic with the features of unconscious emotional states or properties.

This is not to say that attention has not already been paid to just such issues in at least some areas of psychological research. Over a century ago, Weber, Fechner, and others laid the foundations of what was to become known as psychophysics. These investigators were interested in the relation of the individual's sensation or perception of a stimulus and the physical properties of the stimulus. By developing such concepts as that of the "just noticeable difference," Weber found that although systematic relations between sensations and stimuli did exist, they were not of a one-to-one order. Rather, the relative amount, or the relative increment in the stimulus, was systematically related to the increment in the sensation. The conceptual framework of psychophysics thus relies on a mental act of discrimination of differences between stimuli of known (physical) magnitudes.† In some disciplines, the act of making such a discrimination is a relatively simple operation, especially where the investigator relies on "pointer readings" and the like. In psychiatry, it is much more difficult, since if involves, first, the definition of the event or phenomenon and then the act of making the discrimination of difference in kind or amount. This complexity has, in part at least, accounted for the unevenness and the lag in the development of appropriate (i.e., isomorphic) measurement techniques in psychiatry.

The concept of measurement has its origin in the extension and refinement of the more primitive and more general idea of classification. The refinement involved may be such that it makes possible the logical use of numbers to represent certain qualities, objects, or phenomena. More specifically, significant measurement results when the principles of a classification are so formulated that they enable us to relate the arithmetic of numbers to varieties of empirical situations. Such an activity is significant because it is enormously helpful to the formulation of the general laws of science, which systematically relate the data being studied, and thus provide for scientific explanation and prediction. The attainment of this level of knowledge is dependent upon the precision with which relevant differentiations can be made in the data under consideration. Such precise differentiations are functions of measurement.

Initial classifications are based upon the recognition of qualitative differences in ordinary experience. But in order to attain the reliability required of scientific knowledge, it often becomes essential to be able to state the degree of these differences. The several ways of quantifying qualitative differences constitute di-

† For a systematic account of the status of "psychematics, measurement, and psychophysics," see Stevens.
verse modes of measurement. That is, the raw material of observation cannot be readily dealt with in the public, repeatable way characteristic of the scientific study of change without being refined or processed in several ways by the observer(s). Useful observation of changes can rarely be expressed in ways which rely upon unaided report of memory alone. These refinements of qualitative and "naturalistic" observation include coding, measurement, and calculation.\footnote{Most of what is being suggested here is well beyond the scope of a necessarily brief report. But it may not be overly ambitious to attempt, in an illustrative manner, to present briefly some of the fundamental considerations involved in the concept of measurement. But even the most modest undertaking in this respect should be prefaced by two cautions. First, the only alternative to a concise, technical use of special notation in the statement of certain logical principles is a discursive presentation that is likely to subject the reader to various expository defects. Second, current literature dealing with theoretical aspects of measurement warrants the judgment that some notions of classical thinking may now be in a state of flux. The contents of this section on measurement drew most importantly upon the following sources to which the reader is referred for further development: (1) C. W. Churchman and Paul F. Fatio (Eds.), Measurement: Definitions and Theories. Note particularly the articles by Stevens, Churchman, and, for advanced mathematical treatment, Suppes. (2) For a new and very useful treatment, see Warren S. Torgerson, Theory and Methods of Scaling. (3) A rigorous analysis of the characteristics of measurement in general is given by Patrick Suppes and Joseph L. Zimtor, "Basic Measurement Theory," in R. Duncan Luce, Robert R. Bush, and Eugene Galanter (Eds.), Handbook of Mathematical Psychology, Vol. 1. The committee is indebted to Dr. John Tukey, Professor of Mathematics, Princeton University, for his informed guidance in measurement theory.} Coding Coding, or systematic naming, is the application of a language\footnote{We adopt this viewpoint for heuristic purposes in keeping with the logic of our exposition, although we appreciate that it seems at variance with the Whorfian hypothesis as to the manner in which linguistic categories affect both how and what we perceive.} to the observation. While ordinary language is a species of coding, and technical language more obviously so, both will often be less useful than some considerably more systematic, planned code. In all these cases, there are specified rules for passing from the object or system, or property of the system, to the code that is to be used. The rules for making this transition may appear very simple—as in simple naming—but they usually turn out to be less simple when they are made explicit.

The first logical conditions of coding and measurement are those which must be fulfilled by any meaningful classification; the classes constituted must be mutually exclusive and jointly exhaustive. Thus, for example, a classification of emotions, would be fully satisfactory if, and only if, it included all the qualitative differentiations that distinguish emotions of different kinds from one another and no emotion could be put under more than one heading. The success of this simple sort of operation depends upon the adequacy of the definition of emotion. An exhaustive classification, however, need not necessarily identify all possible candidates for inclusion; a classification based, for instance, on (1) feelings of guilt, (2) all other feelings, and (3) the absence of feeling would be exhaustive in this sense.

What follows from these considerations is that any successful measurement of psychological states would first require explicit definitions (verbal or extensive) of the terms used in the related theory. With such definitions questions would never arise, whether a psychological concept was a theoretical construct, an inferred entity, or an observable datum. As Rapaport observed, it is not surprising that "most of the experimenters who have attempted to confirm or refute the relationships postulated by psychoanalytic theory were unaware of the nature of, and the variables involved in, the relationships which they set out to test." The problem of measurement in psychiatric research is often just this one alluded to by Rapaport—the measurement of nonobservables, of inferables. This problem is by no means unique to psychiatric research. In all disciplines, the rules for acceptable measurement demand strict lines of concordance between inferables and observables (for example, between inferences about the behavior of electrons and the discernible phenomena in the Wilson Cloud Chamber). In psychiatric research, these lines are often more diffuse and more difficult to establish.

In general, good classifications are necessary to scientific de-
development; substantial progress in the several sciences has been
the consequence of the successful extension of classifications into
forms of measurement. The establishment of explicit criteria for
placing an item in a category of a certain name may prove both
fairly complex and genuinely important. The confused status of
diagnostic entities in psychiatry is an eloquent example of the
problems encountered in making these rules of naming clear and
explicit.\footnote{In this connection, see Raymond Broek: "Philosophical Aspects of Medical Criteria."\textsuperscript{13}} In many cases, it is most economical to define the
criteria for many items at the same time—in order to facilitate their
inclusion as variables in some coding plan, for example.

There are many reasons for coding. First, coding generally
entails a more or less standardized procedure—a procedure
specified in such a way that it provides a basis for adjusting expe-
riences in widely differing contexts, by different people, or at
different times. In simple cases, standardization may mean that
little or no adjustment is necessary for a different time, place, or
person. At its most sophisticated level, standardization gives di-
rections on how to make adjustment, derived from certain rules
that can be made explicit. Well-planned coding is more general-
izable, and it is also more precise, in that it enables one to distin-
guish objects and their properties with some arbitrarily assigned,
specified degree of refinement—one can say how precise the re-
port is, in contrast to the situation using ordinary language. Fur-
ther, such specification makes it possible for the investigator to
come to an explicit decision about the degree of specification
needed for the task at hand. Too-inclusive a specification (for in-
fstance, of what constitutes anxiety) will often fail to allow
significant discrimination; too narrow, may exceed the limits of
our capacity to discriminate.

All coding, including ordinary language, entails the sacrifice
of some information possessed by the observer. Making this rath-
er banal point is necessary, since the clinical method often seems
to attempt to encompass all characteristics of the person or sys-
tem. Coding and measurement are often criticized as procedures
because of the apparent loss of information or richness that their
use entails. Coding is schematizing, but the question should be
whether the coding is appropriate to the problem at issue, whether
it omits material \textit{significant to the purpose}—not whether it
constitutes a theoretically complete description. This judgment
depends upon the purposes of the specific investigation and can-
not properly be the subject of a general indictment. To say that a
man is six feet tall is to say something reasonably clear and pos-
ibly important; yet this statement is not an exhaustive account of
all his attributes.

The usual coding systems utilize numbers at some point. This
is for a very specific reason. Language is not only used for com-
munication and expression, but also for calculation, and the
usual system of numbers is adapted to this purpose. Numbers
may be used in several different ways. They may be used simply
as names—for instance, for baseball players.\footnote{S. S. Stevens\textsuperscript{4} employs the term \textit{nominal scale} for this usage. Most other authorities
prefer to restrict the term \textit{scale} to operations that include at least the specification of
ordering items in a series.} More important for our purpose, numbers may be used to count—a system of
matching the natural number system with an item-by-item array
of objects. Number is an invariant property of a group. Arith-
metic is one way to introduce precision into many statements,
through the process of enumeration or counting. Counting would
not normally be done simply to avoid vagueness, but because
there was already some idea to which a numerical inventory
would be relevant. An example of such usage in our field is a
symptom checklist like the Minnesota Multiphasic Personality
Inventory (MMPI). This relies on direct counting, with items ei-
ther included or excluded from the count. The use of such a list
facilitates comparison and calculation. This is, however, a kind
of operation significantly different from that involved in the use
of a degree-of-illness scale, which is a \textit{measurement} (to be dis-
cussed in the next section).

All counting presupposes clarity of prior knowledge. Where
adequate definition of what is to be counted is lacking, the pro-
cess of enumeration itself cannot be relied upon to serve any scientific purpose—that is, where the categories themselves are imprecise, the exactness of the count of what has gone into the categories is meaningless. Moreover, counting may only be done within a group of discrete objects, events, or qualities that are distinguishable one from another. It would thus be questionable whether or not one is really counting the frequency of, say, hostile acts or feelings of depression, without the explicit introduction of certain conventions by which it can be decided what to do about vague cases. Counting has been done without the use of such stipulated agreements about meaning, but the process in such cases never solves the problem of having to know what was counted.

While coding an observation often will involve the use of numbers, either in counting or in measurement, it need not always do so. There are many non-numerical codes that are being turned to relevant uses in psychological science in promising ways. Examples are the use of symbolic logic in analyzing decision problems and the use of formal codes in linguistic structure analysis.

Measurement*

Measurement is a highly special form of coding, marked by the assigning of amounts of a property to a scale. The simplest form of measurement, which does not involve the assignment of numbers, is called ordering. Here we may deal with the question, whether there is more or less of some property in one case than in another, without having to answer the question, how much more or less (the hardness of rocks, the ranks in the Army). What results is an ordinal series.

* By way of preliminary clarification, we wish to distinguish here between mathematics, which has quantifying and nonquantifying branches, and measurement, which always involves quantification. However, measurement does not thereby necessarily depend on numbers. Ordinal scales, in which positions are assigned along a dimension of possessing more or less of a given attribute, are measurements without numbers. As already mentioned, the "ranking" of baseball players by numbering them (the nominal scale of S. B. Stevens) is not considered measurement by most authorities.

The extension of a classification into an order logically depends upon the existence of a relationship having the properties of asymmetry and transitivity. Thus, if A is more hostile than B, it must be false that B is more hostile than A (asymmetry); and if A is more hostile than B and B more than C, we must be able to infer that A is more hostile than C (transitivity). The first of these properties, asymmetry, is generally easy to test, but the second, transitivity, may be somewhat more complicated. Because our concepts are often multidimensional rather than unitary, a variety of tests of interrater agreement are often employed, presumably on the assumption that a partial ordering of preferences among trained observers will justify a selection from among alternative hypotheses. But an analysis of such a situation will not always show the relationship of transitivity necessary to a valid ordering.

Any sort of ordering depends upon our ability to discriminate relations of properties that are one-directional (i.e., asymmetrical). An order is said to be simple or complete when it involves the ranking of every member of the selected sample. Thus, all positive integers can be ordered by the relations of "greater than" or "less than." Linear orders are simple. An order is said to be partial when some members of the class of things or properties involved are not comparable. This occurs when the attributes of the relevant set are multidimensional. Perhaps the clearest example of a partial ordering would occur in a ranking of automobile license plates, for those which use letters would not be fully comparable with those using only numbers.†

Ordering can also be primary or secondary. Primary ordering is based on observables—the size of objects or the degree of a response. Secondary ordering is not based on direct observables but is inferable on some basis other than direct observation of the phenomena themselves. For example, colors can be ordered in many ways and, except for the rainbow, have no a priori "nat-

† See R. L. Ackoff, Chapter 6,* from which the above illustration is taken, and which will provide the reader with a more extended exposition of some of the technical differentiations among orders.
natural" ordering. But they can be ordered meaningfully (red, orange, yellow, green, blue, indigo, violet) on the basis of the length of the corresponding light rays. Phenomena can also be ordered in terms of theory. For example, the core neurotic conflicts of a group of individuals can be ordered in accord with psychosocial fixation and regression points as represented in psychoanalytic developmental theory.

For the description of research results in psychosocial science, ordering is at times the only available, and often a satisfactory, means of expression. But powerful theory construction may require more than this. A theory is inadequate unless definite and verifiable consequences can be deduced from it. The likelihood of obtaining these may often depend upon greater degrees of mathematical precision than that embodied in ordinal scales.

Ordinal scales, which represent simply an ordering of units, have no implication for distance between units and no necessary natural point of origin. Interval and higher-type scales embody increasing amounts of this more precise information. Measurements in any science vary greatly in the extent to which one can specify their degree of precision and in the type of computations that can be made from them. Many of these considerations are similarly applicable in psychosocial measurement, and a brief classification of scale types and descriptions of their characteristics will suggest the array of tools currently available.

Scales may be classified (1) according to their internal characteristics, (2) according to rules for their development, and, in several ways (3) according to their place in research strategy. The internal characteristics of scales are order, distance, and origin. Order—that one position on the scale precedes another—characterizes all scales and may be considered as a defining characteristic. (This omits the special use of the term "nominal scale" by S. S. Stevens, mentioned earlier.) Distance refers to whether or not the spaces between points on the scale have specifiable meaning. Origin refers to whether the zero point on the scale is arbitrarily or naturally defined. Thus, according to Torgerson, scales may be classified as follows:

<table>
<thead>
<tr>
<th>Distance has no meaning</th>
<th>No Natural Origin</th>
<th>Natural Origin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ordinal scale</td>
<td>Ordinal scale with natural origin</td>
<td>Ratio scale</td>
</tr>
</tbody>
</table>

In temperature measurement, the first of four levels discussed by Campbell is an ordinal scale. At this level, each measurer makes and calibrates his own individual thermometer, with no necessary correspondence between thermometers except that each implies some directional response to temperature. The second level is an interval scale. At this level, arbitrary dimensions (glass tube of particular bore, mercury as the liquid) and units (setting 100 degrees between the freezing point and the boiling point of water) have been introduced. The third and fourth higher Campbellian levels, which involve defining absolute zero in terms of the properties of nature, are ratio scales. Street addresses on the same side of the street are a more mundane example of an ordinarily ordinal scale, the differences between the successive numbers not being a measure of the distance between the houses, but the sequence corresponding to the sequence of houses. Street addresses may or may not have a "natural" origin. Another example is a foot rule, which is an interval scale, since the zero is set wherever you want to set it.

Scales may also be classified according to the rules for their development, that is, for assigning their relationship to the property being measured. Three types occur: fundamental measurement, derived measurement, and scale assignment by fiat.

In fundamental measurement, each point on the scale refers to an amount of the property measured, so that the scale is a "map" of the property measured and it is possible to show operationally that the relation between two or more points on the scale corresponds to the relation between the two amounts of the property measured. An example is described by Campbell in his treatise—the ability to combine physically two quantities of weighed water and find that the total weight corresponds to the sum of the addition of the numbers denoting the initial weights. It must, of
course, be borne in mind that it is the weights that are additive, not the number of drops. In terms of numbers, two drops of water added together make one. Similarly, the addition of length and of other properties is also possible. Our very familiarity with these measurements often interferes with our perceiving their operational and logical structure. In the psychosocial field, that kind of fundamental measurement does not exist.

Derived measurement refers to indirect measurement, found by combining two other measurements in some specified way according to some theoretical law. For instance, density is derived in its usual form from the ratio of two fundamental measurements, weight and volume. Again, psychological measures cannot be combined in this way. There is no such thing as "adding" the general abilities that are measured by intelligence tests.

In the psychosocial field it is usually necessary to assign scale values to properties arbitrarily (i.e., in the absence of external criteria) rather than operationally. This is called by Torgerson® assignment by fiat. The usual intelligence tests are of this type—that is, one can add intelligence quotients and derive means, and this has meaning, but not the same mathematical meaning as adding the numbers (an IQ of 140 does not signify an intelligence exactly twice as great as an IQ of 70).

Scales may also be classified according to their place in research strategy. Thus they vary according to:

1) **Procedure**—rating, sorting, paired comparisons, rank order, choices.

2) **Focus**—subject scaling (where individual differences between

---

® In fact, the chief significance of the property *number* is that it is increased by addition. This latter consideration is a *Deus ex machina* of the process of measurement and helps to make understandable the assertion that "the actual excellence of physical measurement is entirely a matter of fact." In his classic statement on the subject, the philosopher-scientist N. Campbell has stressed that "... measurement does depend upon experimental laws; that it does depend upon the facts of the external world; and that it is not wholly within our power to determine whether we will or will not measure a certain property. That is the feature of measurement which it is really important to grasp for a proper understanding of science." For a presentation of arguments that are critical of these notions of fundamental measurement see R. L. Ackoff, pp. 192-201.

---

3) **Nature of response**—relation of stimulus to subject is of interest (e.g., sweetness); relation of stimulus to specified attributes is of interest (e.g., attitudes in specified situations, like authoritarianism).

4) **Comparative**—where explicit statement of the relation is given; versus

**Categorical**—where action (sorting, or the like, into classes) is the basis of measurement, without all the defining relations being made explicit.

Two reasons may be adduced for at least listing these various scale types: first, to indicate the array of possibilities available to the investigator; second, to highlight certain crucial aspects of the problem of measurement in general. Much discussion about what is measurable has centered around the previous misuse of scales—for instance, the failure to note that adding or dividing scores on certain types of scales is an entirely meaningless and confusing operation, while lesser calculations could have been of value. The planning and selection of the appropriate measure can and should be done in advance, often with consultation in what is now a highly specialized field. Recognition of the limitations of particular methods of measurement can preclude their misuse and can leave measurement as the servant of analysis rather than its master.

Very significant measurement results when the principles of a classification are so formulated that they enable us to relate procedures characterizing the arithmetic of numbers (like addition or multiplication) to certain varieties of empirical situations. Such an activity is significant because it is enormously helpful, although not absolutely indispensable, to the formulation of the general (invariant) laws of science that validly relate the various

---

© For example, Darwin's theory of evolution by natural selection.
data being studied and thus provide for scientific explanation and prediction. The attainment of this level of knowledge is dependent upon the precision with which relevant differentiations can be made in the data under consideration. Such precise differentiations are functions of measurement.

In discussing measurement possibilities we are faced with the fact that familiarity may breed not only contempt, but also ignorance. The elegance of simple arithmetic calculation is so familiar that its complexity and values were only appreciated many years after it was in common use. While current preoccupation with computers has reawakened most intellectuals to the importance of calculating (even to the point of apprehension), and while we need to caution against the uncritical use of numerical methods, the advantages to be gained for the analysis of data by making it accessible to calculation are so great as to be worth great effort. Calculation is no substitute for analytic thought, but properly applied, it may often do for analytic thought what cannot be done without it.

The complexities involved in measurement in psychiatry should certainly not act as a deterrent to continued interest and the consequent development of new measurement techniques. But it does seem useful for the psychiatric investigator to recognize that certain difficulties might stem from confusions about the nature of measurement and from baseless assumptions making its role in physics the model for every other science. The common allegation that measurement is the essence of science has received serious objection from Kantor,

... measurement is not identical with observation, with research, not even with manipulation. Why confound one type of means with the character or end of science? Worse still, the idolator of measurement confounds the means with the things investigated. ... After long entertaining the dogma that science is measurement and physics, because built upon mensurational operations, is the primary science, workers in quantum mechanics have come to the following conclusions: (1) measurement cannot be regarded as operations performed upon independent things with unique inherent properties, and (2) measurement is the process of creating the things measured.49

Finally, researchers should not overlook the fact that it is an arbitrary restriction upon the meaning of measurement to identify it with the assigning of numbers. Many agree that operations which deal with configuration and pattern are essentially mensurational. Forms of measurement are capable of being derived from topological mathematics that are not necessarily restricted to numbers and that are not necessarily more complex. The development of such means of measurement (e.g., the scattergram configuration on the Wechsler Adult Intelligence Scale) would seem to hold some promise for psychiatry, since so many aspects of psychological relationships do not fit readily into the more traditionally recognized types of quantification. And in speaking of the newer mathematical theory of games, Abraham Kaplan49 observed that "it has shown that the resources of the human mind for the exact understanding of man himself are by no means already catalogued in the existing techniques of the natural sciences."

Problems of Measurement: The Phenomenon as the Variable

We have seen that various scales (nominal, ordinal, interval, and ratio) have been developed to assist in the measurement of phenomena. Concomitant with the use of such scales, the investigator has two tasks: (1) to reduce the phenomena he is studying to data, and (2) to reduce the data to statistics on the basis of which inferences can be drawn and generalizations made about the phenomena. In the first task the investigator makes use of the available and appropriate "instruments" to obtain measures that represent the dimensions of the phenomena with minimal distortion or error. Such instruments may range in objectivity and precision from overall clinical judgments on rating scales to the dial readings of a highly technical apparatus. In the second task the investigator may use a variety of statistical techniques to reduce the measures or data to an appropriate summary form in order to ar-
rivate at general statements regarding the phenomena under speci-
fiable conditions. Maximal congruence should exist between the
properties of the phenomena, the measures, and the statistical

techniques.

Phenomena may vary along several dimensions. For example,
they may vary in the extent to which they are observable, rang-
ing from directly observable (e.g., overt movements) to indirectly
observable (e.g., heartbeat) to inferable on the basis of behav-
ioral events (e.g., anxiety) to nonobservable theoretical con-
structs (e.g., ego strength or anxiety tolerance). Phenomena may
also vary in terms of such other dimensions as their complexity,
extensiveness, or duration. For example, in the univariate case,
where only a single measure or variate is involved, the pheno-
mena may be monotonic and generally nonreversible (e.g., physi-
cal height); may be usually monotonic but partially reversible
(e.g., body weight); or may fluctuate within definable homeostatic
limits (e.g., blood pressure). In the multivariate case, where
several measures or variates are involved, the problems of mea-
surement and data analysis may be considerably more compi-
lcated, depending on the number and relative independence of
the measures. In some cases, the set of measures may act more or
less in unison, or with relatively consistent or stable relations
among the several variables involved (e.g., systematic reactions
to particular stressors, stress tolerance, and recovery from stress);
but in other cases the relations among the variables may alter, for
example, with time (e.g., the affective and cognitive components
of patients' reactions during psychotherapy).

Keeping in mind the various properties of the phenomena he
is studying will help the investigator to determine the optimal
techniques for data collection, data processing (the possible need
to rescale or otherwise transform the raw data) and data analysis
(the possible appropriate statistical techniques) and so maximize
the extent to which his conclusions actually reflect the properties
of the phenomena.

In this discussion, we will touch upon some of the problems
frequently encountered in the attempt to measure change, re-
gardless of whether the investigator is interested in one or per-
haps several concomitant and/or sequential measures of
change. Change is seldom, if ever, measured directly. Rather, it
is implicitly or explicitly indicated by the difference between two
or more separate observations or measures over some period of
time. The investigator may be interested in one of several mea-
sures of change. He may be interested in measuring change on a
single variable in a single individual, or measuring change in a
set or pattern of measures within an individual. This approach is
usually called idiothetic, involving ipsative measures or scales.
Or the investigator may be interested in measures of change over
several individuals, involving one or more measures, which may
be independent but probably will be interrelated, that is, cor-
related. This approach is usually called nomothetic, involving
normative measures and scales. Correspondingly, the investigator
has at his disposal a variety of statistical techniques for analyzing
idiothetic and nomothetic change data. In the idiothetic approach
he may use univariate measures and employ time series, lag
correlations, sequential analyses, and some forms of factor
analysis. In the nomothetic approach he may use simple, partial,
or multiple correlation methods, regression analyses, tests of
significance of differences, factor analysis, or other univariate or
multivariate statistical methods (cf. Bock and Haggard).•

The type of research problem and the type of scale used will
tend to determine the types of statistical techniques the investiga-
tor may use. With a nominal scale, used in the sense of the clas-
sification of individuals into unordered categories in terms of a
particular characteristic, he may measure change by the use of
the Chi-square test for correlated proportions (McNemar,• pp.
228-230). With an ordinal scale, which involves the ranking of
individuals, the investigator may measure the change in ranking
of traits within one person ("although his outstanding quality
used to be hostility, it is now assertiveness"); or a change in the
ranking of persons within a group ("Joe used to be the best in the
class; now Bill is"). Change is most usually indicated, however,
in terms of measures on an interval scale, since the arithmetic
operations of adding, subtracting, multiplying, and dividing can then be readily employed. Perhaps just because the interval scale is so seductively easy to manipulate arithmetically, at least two considerations should be kept in mind when one attempts to measure change on this type of scale.

First of all, the measure may be fallible insofar as it does not measure accurately, or does not measure only (i.e., measures "things" in addition to) what the investigator wants to measure or thinks he is measuring. In such instances the investigator might be quite misled if he were to take at face value the raw or arithmetic difference between two fallible measures on an interval scale, assuming it to be the true, that is, valid, measure of change in the phenomenon. We will consider this possibility in more detail presently. This fallibility problem of course occurs with every type of scale, not just interval scales. But with ordinal scales the investigator will not get into this particular dilemma because he will know that subtraction used in order to find a difference is, to begin with, a meaningless operation with the scale he is using.

Second, it frequently occurs that the scalar properties do not actually correspond to those of the phenomena to be measured—for example, when the phenomena are nonlinear. Since the interval scale is by definition linear, if it is applied directly to nonlinear phenomena and the scale scores are taken at face value, one will be misled into assuming that the phenomena are also linear. This incongruence between the phenomena and the scalar values representing them may go unnoticed when, for example, the scalar units are physically—but not behaviorally—equivalent. Thus, although the difference between 98 to 99 degrees Fahrenheit and 103 to 104 degrees Fahrenheit is in one respect the same (one degree in each case), the "behavioral meaning" in the two instances may be very different and one cannot assume that the behavioral correlates of the temperature changes are linear in the same sense that the temperature itself is. Similarly, growth rate may look linear because it is measured in linear time units, but clearly it is not. The discrepancy between the phenomena and the scalar values also occurs when the phenomena are nonobserv-

able, although in such cases detection of the discrepancy is even more difficult.

We observed in the foregoing that estimates or measures of change rest upon separate observations made at two or more points in time, with change defined in terms of one or more discernible differences between them. Therefore, before considering the problems involved in measuring change, some attention should be given to the measurement of phenomena at a fixed point in time. The theory of measurement which is based on scores pertaining to nonobservables has been developed most highly in such areas as education and psychology. In the classical physical sciences, this type of measurement did not need to be developed to a high level, since the physical scientist was able to take more or less at face value his direct observations or the measures obtained by his measuring instruments. His measurements of height, weight, distance, temperature, and the like were sufficiently accurate for his purposes, so that he did not need a theory of measurement—or elaborate statistical analysis—to help him reduce and evaluate his data. His measures were sufficiently pure or infallible, or they contained sufficiently little error of measurement, that they could be used directly in plotting (e.g., the relation between increased temperature and the expansion of gases) or in formulating his scientific generalizations. After physical scientists began to deal with microphenomena, such as nonobservable particles in the Wilson Cloud Chamber, they were faced with problems of measurement and data analysis that the social and behavioral scientists have dealt with for a long time, including heavy reliance on inferred measures and on statistical techniques.

In the social sciences, and particularly in areas like education, such phenomena as intelligence, reading ability, or other forms of intellectual accomplishment cannot be measured directly and infallibly. This is partly because they are complex functions, but primarily it is because the phenomena to be measured are not directly observable. Under such measurement conditions it is well known that factors other than those the investigator wishes to mea-
sure enter into his measurements. It is to be expected, for example, that degree of motivation, fatigue, attitude toward the investigator, familiarity with the task, among other factors, may determine in part the score an individual will obtain—even though the test is designed to measure intelligence or mathematical ability. Insofar as the various factors that contribute to the ability or achievement test score are irrelevant to the investigator's interests, they are called errors of measurement. One of the main purposes of measurement theory for this type of case involves an appraisal of the magnitude and effect of such errors of measurement.

We may then view any score on such a test as composed of at least two parts or components: the "true" score, which refers to that portion of the obtained score which measures what the test purports to measure, and the "error" portion, which contains the remainder, irrelevant to the true score. Since it is not possible to separate and designate these two portions directly, the usual procedure is to estimate the average expected size of the error portion, and then to subtract it from the obtained score in order to estimate the true score. The variance of the error of measurement can be estimated rather easily if the test (or measure) is repeated on a large number of occasions under the same conditions. If, under such idealized conditions, errors of measurement exist as part of the obtained score, on replication the scores will differ from one another, and hence will form a distribution of obtained scores. The assumption is made that the mean of the distribution of obtained scores is the best estimate of the true score, and that any score which deviates from it reflects measurement error. Usually, it is also assumed that the errors of measurement are distributed normally about the true score, and that the magnitude of the errors will be independent of (i.e., not correlated with) the magnitude of the true score. But if the errors are distributed normally with zero mean, it follows that over a large number of trials, the errors will add to zero—that is, the sum total of the errors will balance out. Given these assumptions, let us suppose that on a particular test (say on trial $n$) the obtained score is far above (or below) the distribution mean. In such a case, the measurement error for that score will be positive (or negative). It follows, then, that on the next trial (say, $n + 1$) the obtained score will be expected to fall below (or above) the preceding score—or, in other words, it will tend to "regress toward the mean" of the distribution of scores.

The phenomenon of regression toward the mean is of central importance in the evaluation of change when the measures contain errors of measurement—and the larger the errors of this type, the more important the regression phenomenon becomes in the estimation of change. In the absence of repeated measures on individuals under the same testing conditions (i.e., with no effects intervening between the testings, which is of course a theoretical but unobtainable ideal) procedures based on the scores of groups of individuals on a single testing have been developed in order to estimate the measurement error of a given test or measure.

Thus far we have considered some of the problems inherent in the measurement of ability or achievement as they pertain to an educational setting. The same set of measurement problems confronts the investigator in a psychiatric setting, especially if he is attempting to measure characteristics such as anxiety. The fact that investigators often have failed to recognize and cope with the problems of measurement as they pertain to their research efforts is one of the reasons why they tend to be disappointed with attempts at cross-validation studies in psychiatric research. It is perhaps disconcerting to recognize that uncontrolled sources of error in the sampling, measurement, and data-analysis phases of the research contribute to the all-too-frequent failure to replicate findings on different samples or in different settings. But one

$\dagger$ If the errors of measurement are distributed with zero mean, so that for any individual his expected error of measurement over many trials (and hence on any given trial) is zero, the errors are called unbiased. If the measurement errors are not distributed with zero mean, they are called biased errors; it is then to be expected that any particular obtained score will likewise be biased. Since such bias can be expected to distort systematically all the obtained scores, the measuring instrument (e.g., the test) should be modified to remove the bias.

$\ddagger$ These are studies designed to discover whether or not the findings of the original study hold up on subsequent trial using a different subject sample.
should be no less disconcerted by the fact that, unless uncontrolled sources of error are eliminated or otherwise properly controlled, the "findings" of any particular research may be more or less invalid—even before they are shown to be unstable by a cross-validation study.§

The phenomenon of regression toward the mean can also occur with essentially infallible direct measures. For instance, in the example of shifts of one degree Fahrenheit at two divergent points on the temperature scale, a regression effect will likewise occur. In this situation the regression effect will result from the fact that the properties of the temperature-regulating system tend to "push" toward a temperature of 98.6 degrees Fahrenheit. Because of the activity of homeostatic mechanisms, if a given bodily temperature reading diverges from 98.6 degrees Fahrenheit on one occasion, it is to be expected that, other things being equal, a reading taken at a later date will tend to "regress" toward this mean value.

It has been considered acceptable by some researchers to apply an interval scale to nonlinear phenomena and to assume that because relatively direct (or objective) measures are involved, the resulting change scores are thereby valid and meaningful. The precariousness of this assumption has been demonstrated in a series of studies of the changes of the level of palmar skin conductance, the galvanic skin response (GSR). In these studies, different scales of measurement applied to the same raw data gave strikingly different (GSR) change scores over different response levels. As a consequence, the same findings may or may not yield data that can be used appropriately with some of the standard statistical techniques and may also yield quite contradictory data, depending on which of several available scales of measurement is used (cf. Haggard43,44). Interpretation of directly presented research findings (and of their statistical analysis) must take account of the more or less hidden complexity actually involved, in order validly to reflect their genuine implications.

How the investigator deals with the regression effect as it influences his measures of change will depend on the type of phenomena he is dealing with and the nature of his measurement devices. If the phenomena are not observable, and hence are measured indirectly, the investigator may have to rely on the statistical estimation of the magnitude of his measurement errors and their consequent effect on the reliability of his change scores. However, if the phenomena are more or less observable, the investigator may be able to determine the particular properties of his measurements and make any corrections (e.g., by appropriate transformations) for their initial nonlinearity. In every case it is important that the investigator pay close attention to the nature of his measurements, their congruence with the phenomena he is studying, and their appropriateness in terms of the statistical techniques he intends to use. For example, nonparametric statistics may be better (more congruent) than high-powered parametric statistics for data that deviate widely from linearity.

One may be interested in the measurement of change from several points of view: the reliability of change, the occurrence of the change, the magnitude of the change, and the correlates of the change. The essence of these considerations is that, especially in psychiatric research, attention must be paid to the nature, the meaning, the appropriate use, the fit, and the limitations of the measuring instruments because we do not yet possess measuring instruments that measure well enough to demonstrate reliably even a reasonable difference between states. It is with the consequences of not having reliable change data that the following paragraphs are concerned.

Reliability of change. The reliability of the measure of change will be a function of the reliability of the initial and final score. Like a clothesline, the more unstable (or unreliable) the initial and final measures, the more unstable (or unreliable) the measure between them—the change measure—will be. Moreover, if the initial and final scores are highly intercorrelated—that is, if they presumably measure the same thing or if the members of a group

§ The effects of measurement errors can be cast in slightly different terms, namely:
The larger such errors, the less reliable the measure, and the more significant the part played by measurement errors in determining the change scores.
keep their same relative positions in the two distributions—the reliability of the measure of change will be much less than the reliability of either the initial or the final scores. This is because the change measure, in addition to its own unreliability, contains within it the unreliability of both the initial and the final measures used to determine it. As a typical example, let a given test or measure be administered twice, yielding a reliability coefficient of .80 on each occasion and with a correlation between the two measures of .70: The reliability of the difference scores will be only .33 if the variance of the two measures is the same (cf. Gulliksen). The measurement of change is also complicated by the fact that, other things being equal, the higher the correlation between measures administered on two or more occasions, the lower the reliability of the difference scores will be. The reliability of the change score may be maximized, but this is done at the expense of the correlation between the initial and the final scores. That is, one secures the maximum reliability of the change score if there is a zero correlation between the initial and the final score. But in that case one may not be measuring the same thing at all at the initial and final points and hence may not be measuring change. In other words, if one tries to maximize the reliability of the change scores under such testing conditions, one may come close to sacrificing the variable on which one is measuring change (cf. Bicerter). In some instances, these difficulties may be alleviated by recasting or transformation of potentially infallible measures before they are subjected to statistical analysis.

The occurrence and magnitude of the change. Because measures are fallible and the traits of the subjects unstable, it may be expected that the regression effect will influence the change scores. Thus, if a strong negative measurement error or trait deviation contributed to the initial score, a less strong negative (or even a positive) error or deviation will be expected to contribute to the final score, even though no intervening "treatment effect" had occurred. Correspondingly, if a strong positive error or deviation occurred on the first trial, one would expect a less strong positive (or even a negative) error or deviation to occur on the second trial. Thus, with no treatment effects at all intervening between the two trials, one would still be likely to find a "gain" in the first example and a "loss" in the second example. Correspondingly, if genuine treatment effects do intervene between the initial and final measures under these conditions, they will be exaggerated or attenuated by the regression effect, and the distortion will be proportional to the magnitude of the measurement errors or trait deviations.

It is for such reasons as this that, other things being equal, experimental treatments assigned to individuals or groups with relatively low initial scores have a built-in bias in favor of a gain, regardless of the actual effectiveness of the intervening treatments—unless appropriate corrections are made. The reverse holds for treatments assigned to individuals or groups who obtain relatively high initial scores. Consequently, if under such conditions the investigator takes at face value the observed (raw) difference between two or more measures, he may conclude (with or without the use of "tests of significance") that no change had occurred when in fact a significant change has occurred, or vice versa. The magnitude of the apparent change, as well as its occurrence, is then a function of such factors as errors of measurement and the regression effect. When there is reason to believe that these factors operate to influence the change scores, appropriate experimental controls or statistical adjustments can be made (cf. Lord).

The correlates of change. The investigator is frequently interested in factors that are related to change scores, and he may be inclined to correlate the observed change scores with the other variables that are of interest to him. But for reasons similar to those we have discussed regarding the unreliability of change scores, he may be misled by such a commonsense and straightforward approach. This is because, when one correlates other variables with a change score, the correlations may be spuriously high or low if the same errors of measurement are present in the change score and in the other variables that are correlated with
it. Appropriate statistical corrections, however, may be made for such potential distortions (cf. Lord\textsuperscript{25}).

In approaching the problem of measuring change, one can think of and examine the various sources that contribute to change scores; or, looked at from the complementary point of view, one can specify the several possible areas of error that may influence the change score. The discussion to this point has focused upon the characteristics of the phenomena being studied—for example, from the point of view of their probable nonlinearity with respect to conventional measuring scales. Difficulty arises here when one confuses the measurement with the phenomenon and pays primary attention to the units on the measuring instrument rather than to the behavioral properties and significance of the phenomena being studied. As already suggested, data may be transformed to make more linear the measurement of the phenomenon or they may be rescaled in other ways to obtain the maximum of information from the observations (Bock\textsuperscript{26}).

In summary, many difficulties in measurement in psychiatric research derive from the use of ordinal scales, to which numbers are assigned indicating only rank position on the scale, numbers that are then treated statistically as though they represented the corresponding equal intervals. This difficulty can be guarded against by adherence to the simplest statistical techniques (based on nonparametric methods) unless it can be demonstrated that the particular data warrant being treated as though they were (close to) linear and therefore may appropriately be arranged along a true interval scale. The usual pitfall here is the tendency to look not at the measure (and its problems) but at the result, and, if one is not aware of and concerned with the fallibility of the measure, simply to accept the result. The widespread introduction of computer aids to the study of data has only compounded this problem.

But there are sources of error other than the characteristics of the phenomena being studied. There are the characteristics of the setting or context and of the subject being studied, and the characterisitics of the investigator and of the instrument—as well as the interactions among all of these. These other sources of error in measurement are the focus of the following sections.

The Setting and the Subject as Variables

In psychiatric research the subject can rarely be regarded as a passive object to be observed and/or manipulated by the investigator. If the subject is a patient in a clinical setting, he is always to some extent a participant in the research. This is most evident when the research deals with his private mental states, his attitudes, emotions, ideas, memories, or other phenomena of consciousness; if the researcher is to obtain this type of information, the patient must be interviewed or questioned and thereby becomes an active participant in the research process.

Similar considerations also arise when the phenomena under investigation are not those of mental functioning or other subjective states. For example, it is almost always necessary to enlist the cooperation of a subject in biological studies, particularly those that involve the collection of bodily fluids, such as blood or 24-hour urine samples; or when objectively observed psychological behaviors are being investigated, such as in research on perception or in studies of social behavior in small groups. In these situations, the purpose of the research usually will be explained to the patient in order to justify the inquiry and enlist his cooperation. That is, in order to obtain information from the subject or even to obtain his assent to the procedures of psychiatric research, some relationship between the investigator and the subject must be established and maintained. The impact of this situation may be posed as the question, "To what extent does the definition of this situation as a research setting influence the phenomena under investigation?" There are a number of aspects to this problem.

A question that arises immediately is, to what extent should the subject be informed about the precise nature of the research? If the research requires that the subject not know he is participating in a research inquiry, some form of concealment becomes
necessary. The investigator needs to be aware of the problems that may be generated by this concealment. For example, what are the psychological—as well as research—consequences of tape-recording interviews without prior approval of the participants? Or, what is the propriety of experiments in which deceptive instructions are given, as frequently occurs in research on the dynamics of small groups?

In these situations, the investigator has to estimate the consequences of his procedures when the research project results are published, or if the concealment is discovered. For example, investigators have been admitted as mental patients to psychiatric wards, without the knowledge of staff or patients, in order to study the patients' social system directly. As one might expect, the subsequent announcement that this research had been going on influenced (often in unpredictable ways) the patients' current psychological status, the nature of their relationship to the hospital, and their continued collaboration in the hospital treatment program—let alone in subsequent research.

The researcher must carefully evaluate the consequences of concealment, regardless of whether it is later discovered or revealed. He must weigh a lesser or a different access to the desired research data against the more comfortable ethical position and the enhanced possibility of long-term continued collaboration, inherent in taking from the start the alternate position of making himself known as an avowedly participating research observer. A research strategy that usually allows no choice but that of concealment is the double-blind drug study. Even here, however, double-blind drug studies have been devised that use two drugs, rather than drug and inert control.

When the subject is informed, different consequences may ensue. Some of these have to do with the impact of defining the situation as research. The mere fact of a research context may itself initiate social and psychological processes that can modify the phenomena under investigation (hence the assessment of change in them). In any interpersonal situation, including the research situation, information and expectations are communicated by implicit and nonverbal cues. The very existence of the research project creates cues that often are unacknowledged. Orris has called these cues “the demand characteristics of the experiment.” Similarly, when the research is conducted in a clinical setting, one may talk about “the therapeutic characteristics or by-products of research.”

Related phenomena have been extensively studied in industrial relations, at least as far back as the studies at the Western Electric Company in Hawthorne, New Jersey. In these studies, it was observed that the mere establishment of research procedures—for example, studies of the effect on the workers' productivity of systematically altering the lighting arrangements—resulted in an improvement in the workers' performance with every change introduced (whether turning the lights up, down, or whatever). The workers' interest in the experiment had greater effect on their work performance than the nature of the illumination. This phenomenon has subsequently been called the “Hawthorne effect.”

In psychiatric research, similar effects often occur in rehabilitation projects or drug treatment programs in mental institutions, particularly with chronically hospitalized schizophrenic patients. The initiation of the research often produces clinical changes not attributable to the specific research intervention, that is, to the group therapy or the drugs. Depending on circumstance, or the meaning to the participants, and on the relationship between them, these changes can be either of clinical improvement or of clinical worsening. Similar examples of the effect of situational or contextual factors include the apparent therapeutic consequences of avowedly diagnostic procedures and the well-known “waiting-list phenomenon.” In the latter, patients may report a significant amelioration of distress after being put on a waiting list for psychotherapy which frequently results in the individual's declining the therapy when it is subsequently offered because it no longer seems needed. This waiting-list phenomenon works in research to vitiate the effort at direct use of the “patient as his own control” in the sense of being able to compare change
during a pretreatment waiting period with change during a comparable later treatment period.

But such effects may go either way. Favored patients selected for research study (and extra attention) may find themselves resentfully shunned by the other patients making up the wider ward society and thus lose a variably important sustaining bulwark.\(^6\)\(^8\) Nor are these effects confined to the behavioral sciences. Similar phenomena have been reported in surgical procedures, as, for example, in the enthusiastic early reports of the successful effects on angina pectoris or ligation of the internal mammary arteries (see the reviews by Beecher\(^54\)\(^55\)). The converse has of course also occurred, in surgery as in psychiatry. Thus, there may be unintended psychotherapeutic effects as well as unintended psychonoxic effects, depending on the nature of the context, procedures, or subjects. Unless this reaction is anticipated the nonspecific effects of introducing the research or the new treatment procedures may be spuriously attributed to the specific innovation. (See Merton’s chapter on, “The Self-Fulfilling Prophecy.”\(^59\))

Often the information to be obtained involves personal data, at times of a highly confidential nature, and a special relationship must be established with the subject in order to insure his willingness to participate. For certain types of research, only a psychotherapeutic relationship will suffice. Unless the therapist and the patient are both convinced of the therapeutic intent and the intellectual integrity of the investigator, certain types of information can probably never be obtained, or if obtained, cannot be used. These considerations are illustrative of what is probably a general rule in psychiatric research: the quality of information obtained is dependent upon the nature of the interpersonal relationship between the subject and the researcher.

The subject (and the investigator) may likewise be responding to aspects of the total situation other than those intended, or other than those of which the investigator is aware. Idiosyncratic meanings may be attributed to the situation, having to do with various events in the individual’s life. For example, the widespread preoccupation that followed the death of Franklin D. Roosevelt or John F. Kennedy, or that occurs during periods of acute international tension, influences importantly the mental activity of the participants—and hence affects the research findings that involve this mental activity. Richard Sterba,\(^7\) using the reports of psychoanalytic patients’ dreams on the night after Roosevelt’s death, illustrated convincingly the major repercussions with respect to the patients’ object relationships to their fathers, their feelings of guilt and depression, and their concerns with death—in each instance, of course, idiosyncratically elaborated in terms of the individual character structure and neurotic conflict situation.

We can illustrate these complexities involved in dealing with the effects of situational factors by considering a psychiatric research ward in which psychophysiological or biochemical studies are being conducted. Frequently, such research is designed to correlate changes in behaviors and/or in personality variables with biochemical or physiological processes. The model most often applied in organizing the ward for such research derives from metabolic wards in Internal medicine or pediatrics. In these settings, the patients are usually regarded primarily as the carrier of some unusual disease process or the source of biological fluids, as urine or blood. The research design seldom includes consideration of the interpersonal relationships on the ward or the psychological meanings of the ward situation to the subjects. The ward is conceived as providing a semicustodial setting in which the patient resides between experiments or for the collection of his bodily fluids. This is not to say that the physician in his dealings with his patients will not be sensitive to such relationships and attitudinal phenomena and will not deal with them in his daily clinical care. However, he frequently does not define them as relevant variables in his research design. That is, as clinician, he tends to view his patient humanistically and holistically; as researcher he tends to view the same patient “scientifically” and atomistically, thereby ruling out of his purview the range of psychosocial variables that may in some instances have a critical
impact on the very phenomena that are the specific object of his interest. In psychophysiological research, where behavior and mental states are crucial variables, such an approach is clearly inadequate.

A number of research studies have documented the ways in which subtle social-psychological forces operate decisively in such research settings. In particular, the total ward situation itself has considerable meaning both for the patients and for the members of the research team (see Renée Fox, Experiment Perilous,98 for a detailed discussion of this phenomenon). Unless the investigator is aware of the possible impact of these social-psychological processes, fluctuations in biological functions may be obscured and spurious significance may be attached to phenomena that are more directly related to the psychosocial context.

A striking example of the influence of interaction between setting and subject upon not only the life functioning but even on the life itself of those moving within the setting occurred when the patients on a geriatric ward at the Topeka State Hospital had to be moved to new quarters after their ward had been severely fire-damaged.99 When the patients were thus abruptly torn from their accustomed setting and routines to a new but otherwise substantially similar ward setting, there was a sharp and otherwise totally unexplained rise in the expected death rate. Within a month after the fire, five patients died and three more died within the next two months, or a total of 20 per cent of the ward population of 40 (as compared with a 7.5 per cent death rate in the preceding three months among these patients). This rise seemed unrelated to any directly deleterious effects from the fire itself. The patients had been speedily and efficiently evacuated. No one seemed seriously injured. Indeed, the hospital personnel had congratulated themselves on having successfully averted what could have been a major disaster.

For many years problems of situational context and its meaning have been recognized in research in clinical pharmacology. Whenever research efforts have been made to evaluate drugs, not only with psychiatric patients but even in cardiac disease or in cancer chemotherapy, the need has arisen to provide appropriate control measures. At first, single-blind techniques were employed to control for the novelty of the situation and for the wishes and apprehensions of the patients. More recently double-blind techniques have been introduced to control further for the possible biases of the investigator and the interacting effects of these biases upon the expectations of the patient. The development of blind techniques and the use of placebo as control have developed into a standard research method in clinical pharmacology. (Cf. GAP Report No. 42, Controls in Psychiatric Research.)

Recently, explicit attention has been drawn to this role of milieu and attitudinal influences upon research procedures and outcomes. For example, it was early observed that reports on the efficacy of the tranquilizer drugs were more favorable from public institutions (state hospitals) than from private hospitals.100 Feldman101 and others have drawn attention to the important influence of the physician's attitude toward the use of the tranquilizer drugs upon the results achieved with them. It was felt that the introduction of double-blind procedures into psychopharmacology might eliminate such effects. Reports of controlled drug experiments usually are less enthusiastic about the efficacy of the drug than reports of uncontrolled experiments. Before it can be assumed that this discrepancy strips away inflated claims and reveals the true measure of the drug's efficacy, other possibilities have also to be considered: that the investigator who conducts double-blind studies is, to begin with, more skeptical about the drug's efficacy and may convey his skepticism, thereby reducing the drug's effects below its true measure; or that in explicitly defining the situation as research, not therapy, the investigator is inevitably behaving with more caution (and with less therapeutic effect) in his relations toward his patients. This hypothesis has recently been put to experimental test in several studies.102-99 Thus, the influence of social-psychological aspects of the experiment as well as of the milieu has become a cogent issue in psychopharmacology. This concern has been widely extended to the
varieties of research in clinical situations, to research in nonclinical human situations, and even to animal studies, where the effect of group processes, as aggregation and previous experience, are now being scrutinized in their relation to drug response.

How can one deal with these contextual variables? We would suggest a number of ways:

First, the nature of the phenomena under discussion must be made explicit. It needs to be accepted that these factors operate constantly and that their effect upon the subject of investigation cannot be eliminated. At the same time, the magnitude of this effect may vary in specific research contexts as slight as to be essentially irrelevant. In each instance, the investigator must make an explicit decision as to whether or not these contextual variables are likely to be of sufficient saliency and magnitude to require inclusion in the research design.

Second, the subject himself may serve as a valuable source of just these data. He may report his feelings, thoughts, and expectations about the procedures before, during, or after the experiment. Information thus obtained from the subject can provide a partial control on whether or not the intended experimental conditions were in fact those perceived and acted upon by the experimental subject.

Third, the role of attitudinal and situational variables has to be explicitly elaborated. Attempts must be directed by the researchers at insuring uniformity of instructions, of responses, and of behaviors. In some research areas—for example studies of perception—this aspect has itself become the focus of intensive investigations. Studies have been undertaken in which the subjects’ attitudes and expectations have not only been assessed but also experimentally manipulated to ascertain the magnitude of their effects on the phenomena being investigated.

Fourth, and most crucial, is the importance of proper control groups. One of the most difficult problems in contemporary psychiatric research is deciding upon the appropriate controls for each research situation or experiment (cf. GAP Report No. 42, Controls in Psychiatric Research). Multiple variables are usually operating complexly, and the research situation in its contextual qualities is a significant part of this matrix of variables that exerts a potent, often multifaceted, influence upon the behavior of subjects studied within it. These facts underscore the need for adequate control by whatever ways are individually most appropriate.

The Investigator and the Instrument as Variables

In all research, the investigator is the final measuring or evaluating instrument, whatever other instruments he may employ. The differences between research in the various sciences and research in specific areas within these sciences can be seen in terms of (1) the frequency with which instruments other than the investigator enter into the making of observations and the gathering of data; and (2) the extent to which assessments are made by the investigator directly, as compared to those made indirectly through the use of other instruments. In psychiatric research the investigator is often the sole appropriate measuring instrument. Much data of vital interest is private and subjective and can only be approached by psychological means. Thus the “instruments” are often clinical judgments of psychological states and the investigator's role as the final evaluator of these elusive data is a centrally determining one. Even in those psychiatric researches that employ other instruments, the investigator operates centrally in the role of interpreter of the meaning and significance of the instrumental readings.

Either the investigator or the instrument (as variables) may introduce systematic, random, or haphazard errors that can occur in the measurements, in the interpretations of the measurements, or in both. These errors may be due to (1) the design and construction of the instrument, (2) the investigator's technique, (3) the effect of the instrument or the investigator on the observed, and (4) investigator subjectivity and bias. These categories are not sharply delimited each from the other; nonethe-
less we shall discuss them sequentially. Our point of departure in each category will be observations reported from laboratory settings.

We will consider examples drawn from general medical and physical science, as well as from psychiatric research areas, to illustrate the variabilities associated with both investigator and instrument. While various studies will be mentioned, two will be referred to more often than any others. In the first of these (from biochemical research), the instrument was the Haldane gas apparatus, the experimenter its operator. The investigators observed the performance of the gas analyses and critically analyzed the procedures and the results; they concerned themselves with the reliability (or variability) of both instrument and operator. The second example is a psychophysiological investigation of the correlation of gastric secretions with psychological processes, particularly affective states, in an infant with a gastric fistula. Here the investigators made systematic observations of the experimenter's interaction with the infant. Thus, it was taken for granted that the experimenter's behavior was part of the data and did influence the experiment. In the first example the research question could be put as "Did the experimenter (the apparatus operator) influence the data, the calibrations, and the determinations?" In the second it could be put as "How and when did the experimenter influence the data?"

Variability due to Design and Construction of the Instrument

Scientists have become increasingly aware of variability and error traceable to mechanical shortcoming within the instrument itself. Dwight Ingle has pointed out some of the possible sources of error in mechanical systems, ranging from the very simple (e.g., the calibration of a hypodermic syringe) to the quite complex (e.g., the construction of apparatus used in the application of tracer techniques to biologic problems). In their study on the Haldane gas analysis apparatus, Renbourn and Ellison found that differences in calibration and in line thickness accounted for some, quantitatively minor, error.

In psychiatric research, the design of measuring instruments, for instance behavior rating scales, presents much greater difficulties. In the study of the infant with the gastric fistula, the investigators established behavioral criteria for the rating of six affects. The criteria were based on descriptions of the child's overt behavior and efforts were made to distinguish each affective state from the others as precisely as possible. However, there was overlap in the behavioral criteria, particularly for affect categories that represented different intensities of the same state—for example, "contentment" and "joy" as two degrees of pleasure. Furthermore, because of conceptual and/or semantic limitations in delineation of the behavioral categories, qualitatively different affects—such as "irritation" (a anger response) and "depression-unpleasure" (a mixture of depression and anxiety)—could not always be reliably differentiated. At times this led to disagreement between two independent raters. Thus, the design of the instrument yielded some degree of precision in rating but also had limitations in this respect. The difficulties in designing behavior rating scales are compounded when groups of subjects are to be observed. Either the rating scale designed for a group of subjects has to cover a very broad (and perhaps too-encompassing) spectrum of psychological processes, or more precisely applicable rating scales have to be designed separately for each research subject.

Some instruments may measure different things at different times. Bereiter cites this example from the domain of more or less standardized psychological testing:

A test that measures arithmetic reasoning ability in children may measure only computational accuracy in the same children when they are older... The person showing high change in the first example will likely be one who was very careful in his computations when older. The bright but careless person could be expected to show less than the average amount of change.

That is, the instrument itself has changed. What had been a rea-
soning task has become a computational task, and the change score may largely reflect one’s differing capacities in these two areas.

Variability due to the Investigator’s Technique

While the demarcation is often indistinct between this category and some of the other categories of variability, that which is based on the manner of use of instruments may be categorized best under this heading. For example, in a study of the arterial blood pressure in a large population in Norway’s the investigators found a bimodal distribution of diastolic blood pressure readings in both sexes and in all age groups. This distribution could not be attributed to social-environmental conditions. The investigators thought that it was most likely due to the observers’ technique in reading the sphygmomanometer at the diastolic level.

In psychiatric research the experimenter’s problems of techniques are great indeed, even when he has to “read” a “well-calibrated” instrument, like a rating scale, with consistency and accuracy. In the study of the gastric fistula the two investigators tried to keep in mind all relevant observational criteria as they rated the infant’s behavior and affect. For the most part, they succeeded in classifying the different behavioral dimensions in a consistent way. However, errors in the proper application of the criteria became obvious when disagreement between the two raters was traced to the “faulty technique” of one of them—his failure to take into account one of the criteria clearly described in the behavioral protocols. Such an error in technique came to light because of a disagreement between raters. There was no mechanism in the research design for the detection of such errors if made by both raters at the same time.

It might be argued that it is the experimenter’s subjectivity and bias that lead to these technical errors. On the other hand, aspects of technique can stand out more clearly than those relating to possible bias; and it seems more appropriate to classify such instances of error accordingly. Errors due to faulty technique may be based on reversible “ignorance” and may be remedial by training. For example, the investigator may not have received adequate training in the precise and sophisticated use of the instrument, or he may come to his tasks with inadequate prior experience in the field of child observation. To use in this context a distinction emphasized by Tukey, appropriate means (proper prerequisite knowledge, adequate training, devices for cross-checking and controlling) can be devised to eliminate or minimize blunder, while appraising error.

Under this same heading of variability due to the investigator’s technique is the additional variability based upon the clinical investigator’s very personal interviewing style and the consequent differing kinds of contact achieved with the research subject. Differences based on such factors of style that are consistent attributes of the investigator as a person (meticulous for detail or freewheeling and encompassing, deadly serious or good-humored, permissive or authoritative) all introduce major variability, especially when the research instrument being used is the interviewing technique of the investigator himself. Anna Freud has given an extended treatment of just these issues in a therapeutic (rather than research) context.

Variability due to the Effect of the Investigator or the Instrument on the Observed

The effect of the investigator’s or instrument’s presence on the state or function of the observed has been increasingly recognized in all sciences, including physical science. While the instrument is the sine qua non for the making of observations, its presence often has a significant effect on the observed that needs to

---

This distinction is between error as the kind of inherent (and inevitable) distortion built into observation by the nature of what is observed, of the apparatus used for the observation, or of what is built into the perceivers in his perceiving predisposition; and blunder as the kind of avoidable mistake that arises from misuse of a technique or procedure and is not allowed by the measuring process. For example, surveying is a science that, through the precision of its checks on all measurements, is able to catch any blunder; however, it can do no better than try to average out the errors, since it is no more or less safe from (systematic) error than is any other scientific activity.
be taken into account. The nature and magnitude of this effect will be a function of the degree of active interference of the investigator or the instrument with the naturally unfolding state of the observed. To the extent that the investigator is aware of such interferences, he will be able to encompass them within his observational framework, and thus turn them to the service of enhanced explanatory power.

In the field of clinical psychiatric research a common problem in this regard arises in the study of the psychotherapeutic process. Certain kinds of data are recorded more faithfully, and are available for continued or additional study more readily, if the therapeutic sessions are recorded on audio or video tape. Sharp debate has occurred within the clinical field about whether intrusion of this new instrument into the usual privacy and confidentiality of the dyadic psychotherapeutic relationship so alters the therapeutic situation that it is no longer the same thing or even therapeutic at all. The fact that there has been debate represents a shift from what Rainy called the “eggshell era” of the preceding twenty-five years—the period of belief that almost any research, by whatever method, into psychotherapeutic processes would of necessity be hurtful. This he described as follows:

Almost without exception, psychotherapists have adopted the attitude that patients and clients were frail, puny beings who would flee the field if anyone (or anything) except their own private psychotherapist touched them. In view of the lack of evidence to the contrary, such an attitude was at one time entirely justified, since, in simple ethics, the welfare of the patient does come first.9

Today, within the framework of the ethical concern of all reputable clinical researchers to do no harm, there is enough evidence to the contrary to have lifted this issue to the realm of an empirical question, not to be prejudged. In practice, of course, there are many difficulties in the way of such empirical study. Because of the irreversibility inherent in accumulated experience, the taping of therapy sessions is not something that can be introduced as a procedure and then taken out, with restoration of the situation to what it was before. Nonetheless, psychotherapy—and, above all, psychoanalysis—gives the very best access obtainable to the meaning, the impact, and the alteration in the process caused by such a systematically added dimension (that can then be encompassed as itself part of the system of interacting variables). The question can thus be shifted from “Does it make a difference?” or more fearfully put, “Does it make a hazardous difference?” to “How much difference does it (necessarily) make, and in what direction, under what circumstances, and with what kinds of patients?” (Similar questions may be asked of the difference it also necessarily makes for the therapist.)

The various studies of placebo effects and the double-blind techniques in psychopharmacological research represent other kinds of efforts to take this phenomenon into account. Even in the physical sciences there are comparable effects but they are usually less significant. Thus, immersing a thermometer in a beaker of water to measure its temperature itself alters that temperature minutely (unless by chance the thermometer and the water had been, to begin with, at exactly the same temperature). But this effect is so minute as to be, with the usual degree of error inherent in clinical thermometers as measuring instruments, negligible. And in those instances where this effect is significant, it can usually be readily calculated.†

In the psychophysiological study of gastric functioning that we have been using as an example, the mere presence of the gastric tube and its repeated insertion and withdrawal very likely significantly modified the gastric secretion of the infant. Furthermore, in order to make the necessary biochemical tests and calculations, the total secretory product had to be removed from the stomach. This removal, too, may have produced significant

†An example frequently used to illustrate this point is that of the Indeterminacy Principle of Heisenberg, the inability to measure simultaneously the precise position and velocity of a microparticle. The measurement of either alters and renders imposable the measurement of the other. This example from the phenomena of microphysics is not directly applicable to the world of macrophysics (and of macro-phenomena in general) with its differing rules, where this kind of effect of the measuring instrument upon what is measured is at least in theory more controllable and more precisely measurable.
changes in the gastric secretory function. The investigators were unable to assess the nature and direction of these changes because there are as yet no reliable instruments for determining gastric secretory rates without aspirating the entire secretory output.

In the behavioral-psychologic area, the investigators in this study took cognizance of the experimenter's effect on the infant and included in their observations that of the interaction between subject and experimenter. Furthermore, the chance observation that the infant reacted by depression and withdrawal when alone and confronted by a stranger was utilized for purposes of study of this new phenomenon by exposing her repeatedly to individuals who were strangers to her. Thus, psychophysioologic data were obtained on the varying effects of individuals having different relationships with the infant.

In this same study observations were also made of unplanned, spontaneous changes in the emotional states of individual investigators and the effects of these on the subject and on the outcome of the experiment. To cite one instance, in the course of one session the experimenter and the observer behind the one-way screen disagreed as to whether to continue with the aspiration. The observer wanted it continued and prevailed. The experimenter, somewhat disgruntled, continued aspirating the gastric juice; he was unsympathetic, businesslike, and impersonal. This seemed to have a marked effect on the child, who responded with considerable anger and an outpouring of gastric juice of high acidity.

In this particular research endeavor, assessing the impact of the investigator as a variable was facilitated by a number of factors: (1) the investigators' awareness of the existence and the operation of this variable; (2) their freedom to observe its operation in themselves and their capacity to deal insightfully with these observations; (3) the systematic use of an observer as a constant aspect of the experimental design; and (4) the opportunity to observe a single subject comprehensively and longitudinally. With regard to this last factor, the difficulties in assessing the experimenter's or the instrument's effect on the observed and on the observations will be different and may become greater as the number of observed subjects increases. Conversely under certain circumstances, the use of the many may control for the variability in observations on the one.

In recent years there seems to have been a shift in the kind of concern felt over this issue. The concern is less now that the phenomena under study may be uncontrollably or unrecognizably altered, and more that the alteration itself should be explicitly focused upon as a systematic dimension added to the research situation, the cogency and impact of which is one of the specific objects of study. In psychophysiological research, investigators thus may now deal with the problem by maintaining the psychologic recording constant and regarding it as a given in the situation. Its influence becomes one of the variables under inquiry, to be dealt with through internal controls—for example, by repeated measures in the same subject. Modern technical innovations (microchemical techniques, telemetry) also make it possible to intrude less grossly into observation and data gathering.

Variability due to Investigator
Subjectivity and Bias

In the course of investigative work, preconception, predilection, and bias in the investigator may all interfere with his perceptions or his thought processes. When an investigator looks at an object in order to find new information, he necessarily sees it within the conceptual framework set by much old stored information, ranging from the very recent to the distant past, ranging from objects almost identical with the one under scrutiny to others only vaguely similar. He will attempt, partly or largely unconsciously, to match the new experience with previous similar experiences. Different observers will tend in varying degrees to see the new in terms of the familiar, in terms of the expected. A striking example of incorrect perception because of expectation of the familiar is given by Johnson Abercrombie in

---

*We are focusing here solely on this particular phenomenon in the investigator. The opposite phenomenon also exists, if more rarely: to see mainly what is different in the new experience.*
her book THE ANATOMY OF JUDGMENT. In the Three Triangles (Fig. 1) most observers fail at first to see the unexpected. As Abercrombie points out, the observer reads the statements as he expects them to be and may need urging or direction to see the extra words. She states further: "When the observer was in State 1 a selection was made so that only those words that form the familiar phrases were seen; it is as though he was blind to the words which did not form part of this pattern, or as though he judged them to be irrelevant background. In State 2 he interprets the pattern on the retina in a different way."

An investigator, in his research, will move from State 1 to State 2 in his own personal way, via what has been referred to in other contexts as his perceptual style or his cognitive style. This style reflects the unique manner in which the adaptive properties of the perceptual and cognitive systems function in different individuals. It even molds and disciplines drive-determined fluctuating motivational states.

Such factors as perceptual and cognitive style will tend to leave relatively fixed imprints on the investigator's ways of perceiving and organizing experience. Thus they may not create much variability in the investigator's perceptions and apperceptions during the course of a research study. Nevertheless, a relatively fixed style of approaching and organizing experience may be a serious source of variability (or of error) if it differs appreciably from that of other experimenters. The study by Renbourn and Ellison demonstrated that different workers have individual preferences for estimating the last value on the Haldane burette: one worker read .3 or .7 much more often than expected and others showed preferences for other decimals. This may be due to one person's tendency to divide a whole into thirds, another into fourths, and so on. However, there are other tenable explanations and much systematic study will be required to elucidate the various factors that may account for relatively fixed differences in perceptual tendency from investigator to investigator.

Of even greater importance to our main theme are the changes in ways of perceiving and organizing experience that may occur in the experimenter over the short range—hours, days, or perhaps weeks. Such changes most likely are consequences of the investigator's current life experiences, be they events within the study itself, in the other aspects of his professional life, or, quite apart, in his private life. A study demonstrating the effect of immediate experience upon perception was reported by Davis and Sinha. Volunteer subjects were shown a painting—Breughel's "Peasant Wedding"—and were asked to describe it. One group of subjects had been told a story of a violent feud between the families of the bride and the bridegroom prior to the wedding; a group of control subjects viewed the picture without having been given this "information." The experimental group perceived the general atmosphere and many details in the painting in a manner quite different from that of the control subjects, who had not heard the contrived story; they emphasized those features of the picture that to them seemed connected with the story of strife, while paying little attention to other, more innocuous features stressed by the control subjects. These last two examples illustrate, respectively, relatively fixed differences between observers due to genic factors and/or past experiences with lasting effects on perception, and presumably temporary differences in perception between observers due to current experience.
Another source of variability is that of differences in perception within one observer, over time or over instances. For example, in one study six highly experienced radiologists and chest specialists were asked to assess changes in pulmonary tuberculous lesions by examining serial chest x-rays in a series of patients, two films having been taken of each patient at three-month intervals. In judging pairs of films with regard to progression, regression, or stability, "a single reader was likely to contradict himself in approximately one-fifth of the cases." The investigators stress the need to determine the factors that influence such errors in film interpretation but did not study them beyond the technical quality of the films, which they did not find to be an important source of error. Other possibilities were not discussed in their report.

In most of the cited examples did the observers’ motivations in terms of manifest wish and need fulfill the perceptual and cognitive process. Yet it is a psychological commonplace that motivation in this sense can have a decisive effect on what is (and is not) perceived and on how these perceptions are organized and understood. The following reported observations will serve as examples. Renbourn and Ellison reported that the quantitatively greatest errors in gas analyses occurred because the workers tended to find the expected results when making the final readings on the burette. The sense of the word expected differs materially here from that used in connection with the Three Triangles. There it meant the familiar, the previously experienced; here it means: a measurement should or will give a result that is in accordance with the measurements of other observers.

An example of the effect of experimenter motivation (or bias) on the organization of thought processes, leading to the systematic exclusion of observed data, was cited by Kety from the area of experimental physics. The phenomenon was the clustering of results obtained in measuring the velocity of light over a 30-year span (1926-1956). Kety stated:

Now even in the measurement of light velocity all the observations are not given equal weight. Rather it is the tradition to reject certain data on the basis of atmospheric or instrumental vicissitudes, although the number of rejected observations is reported. It is my hunch that in this selection process there is an unconscious bias in the direction of conformity with preconceived notions and accepted values. This hunch finds support in the opinion of an authority like Prof. J. A. Bearden, who discussed this question in the American Journal of Physics (Bearden and Tomesen: Am. J. Physics, 27: 569, 1959). After considering and arguing against pure chance and systematic error, he suggests "that the experiments were not really independent but that there was a subconscious psychological factor which tended to make each experimenter look for errors in technique until he could check the then accepted value. It appears that (this assumption) is the most plausible."

Mulligan and McDonald have also discussed this point. "It does seem clear from the history of physics that occasionally the result of a very high precision measurement of a physical constant by someone eminent in the field has intimidated other workers from publishing results in substantial disagreement with this value. . . . As early as 1947 Aslakson found that his 'shoran' distance measurements required a value of c higher by more than 15 km per sec than the then accepted value. He did not report his results at that time, however, because he thought there must be some as yet undetected error in his work or in the accepted value of the index of refraction at the frequency used which would account for the discrepancy. It was only in 1949, when evidence was mounting from other sources in favor of a higher value of c, that Aslakson first published his original evaluation, c = 299,792.4 km per sec. It seems at least possible that other experimenters in the years from 1934 to 1949 may have found higher values of speed by optical methods but refrained from publishing their results because of their disagreement with the determinations of (others) in which such great confidence was placed at that time" (J. F. Mulligan and D. P. McDonald: Am. J. Physics, 25:180, 1957).
In the end, no matter how “scientific” the discipline, there is a question of judgment rather than of empirical observation at issue: When is the weight of new (and different) measurement to be taken over the weight of authority? In physics, the objectivity, sensitivity, and accuracy of the instruments make it likely that sooner or later errors such as those cited by Kety will be corrected. In biology, the prolonged duration of the effects of experimenters’ bias may be seen in the politically motivated concepts of Lysenkoism that held sway for more than 30 years in the Soviet Union and were presumably supported by data in plant breeding. In the last decade, with the rapid rise of molecular biology (and the changing internal political climate) Lysenkoism has ebbed. New data consistent with Western genetic concepts are being obtained, and the reasons for the earlier bias are being examined.

In fields of inquiry where the data are necessarily “softer” and less complete, it is likely that motivational bias exerts an even greater influence. In part this judgment of motivational bias itself depends on the special vantage point and the partisan interest, that is, the bias of the judge. A case in point is the rancorous controversy over the cogency of Eysenck’s publication of findings which affirm that there is no established evidence of the efficacy of any psychotherapeutic approach over nontreatment in the amelioration or cure of psychiatric illness. To Eysenck’s advocates the evidence cited in support of his conclusions is convincing; to his opponents it appears that there are such major methodological and logical flaws in the argument that its conclusions can be properly discounted. Opponents in fact are convinced that Eysenck has demonstrated mainly his own antipsychotherapeutic motivational bias, which resulted in his stacking the cards unfairly in order to make a debater’s, rather than a scientist’s, point.

These considerations of investigator subjectivity (preconception, predilection, bias) are frequently so crucial in what the investigator perceives of the phenomenal world that, without realizing it, the observer may merely rediscover his idiosyncratic way of ordering reality or the way he expects or wishes it to be. The various obstacles in the way of objective scientific scrutiny have been highlighted in various disciplines, in ways depending on the conceptual framework of the discipline, such as the personal equation, countertransference, experimenter bias, and the like. The common denominator touched by these terms is an aspect of this systematic error due to observer distortion or bias.

The “personal equation” has been widely discussed in the literature of the natural sciences. Two definitions from standard dictionaries in the field of physics are the following:

In making measurements of any character every observer, no matter how skilled he may be, is bound to make certain errors. These errors are of two kinds: accidental errors which will be small in the case of a good observer, and which will be distributed in accordance with the laws of probability; and systematic errors or errors which are always in the same direction and of approximately the same magnitude. The value of the systematic error is known as the personal equation of the observer. Personal equation must be determined empirically for each observer under a variety of different observing conditions. For a good observer, the personal equation remains remarkably constant over long periods of time and may be applied directly to any observation. (International Dictionary of Physics and Electronics)

Many measuring operations involve a sensory judgment by a human observer. He may have to read the position of a pointer on a scale; or set the crosshairs of microscope or telescope in coincidence with some suitable target; or adjust two parts of his field of view to equality of brightness or colour; or adjust two sounds to equality of loudness of pitch; or in general produce some specified relation among the objects under observation. Underlying all such procedure is the assumption, broadly justified by accumulated experience: that there is some unique criterion in the subjective structure of the observer’s sensations corresponding to the specified objective relation, . . .

Early in the history of measurement it was discovered that the sensory criteria corresponding to prescribed external relations are subject not merely to random errors within the range of imprecision but to systematic error, the individual errors being more
on one side of the true value than the other, so that the mean of a
series tends towards a value differing significantly from the
truth.

It is this systematic error inherent in the observer's sensory ap-
paratus, as distinct from the systematic errors inherent in the ex-
ternal equipment, which is signified by personal equation. . . .
Nowadays the term personal error . . . is more often used. (EN-
CYCLOPEDIA DICTIONARY OF PHYSICS)\textsuperscript{85}

Kety's discussion\textsuperscript{86} of the systematic distorting factors opera-
tive in the making and reporting of measurements on the velocity
of light is a specific example of the working of this personal
equation. Similarly, in psychological science, in the study of drug
effects, considerable attention has been given to the differences in
ratings between the patient's (involved) personal psychiatrist and
the presumably more impersonal research team. Investigations of
the effect of experimenter bias on the results of laboratory re-
search with humans and also even with animals\textsuperscript{85-88} are relevant
to these same issues. Rosenthal and Lawson,\textsuperscript{88} for example, dem-
onstrated that when experimenters believed that their subject rats
had been bred for brightness, their rats performed significantly
better in almost every experiment than did rats drawn randomly
from the same colony but handled by experimenters who were
led to believe that their rats had been bred for dullness. The only
difference between the experimental conditions were the exper-
imenters' beliefs about their subjects' (rats) abilities. These
differences were obtained despite the fact that the instructors
gave more help to experimenters whose rats were performing
poorly and that all the experimenters were motivated toward hav-
ing their rats perform well in order to complete the sequence of
experiments.

In the clinical fields of psychoanalysis and psychotherapy
there has been a voluminous and ever-growing literature on the
countertransference\textsuperscript{*} the direct relevance of which to these same
issues need only be stated. In fact, with suitable modification,
most of this clinical literature on the countertransference, in all
its range and complexity, is directly applicable to the research
issues considered here. Different linguistic uses and conventions
may likewise lead to different evaluations of what is observed.
This is a thesis which has been most explicitly formulated in the
Whorfian hypothesis,\textsuperscript{89} that one's view of reality is directly molded
by one's language, particularly in its syntactical and lexical fea-
tures.

These many kinds of observer error are not necessarily mu-
truly exclusive. They may act independently or in concert. In
fact, any combination of these mediating avenues may be instru-
mental in leading to an erroneous scientific conclusion. It is also
important to emphasize that these varieties of observer error are
not simply the evidence of wish-fulfilling tendencies—or that
human propensity to find exactly what one is looking for—but
are much more broadly a panorama of the many dimensions of
the truth that what one finds and reports is determined by the
ways in which one looks and thinks, and with what instrumental
and language tools one operates. These observer errors can be
minimized only by conscious foreknowledge that they do occur
and that steps need systematically be made to take their influence
into account.\textsuperscript{†}

There are also many external factors that may heighten an in-
vestigator's motivation for finding the expected, for conforming
with preconceived ideas. Among them are personal loyalties—or
more crassly, coercive expectations—in relation to the depart-
ment head, esteemed teachers, or colleagues in the institution; or
more broadly, the expectations of the larger professional world,
the school of thought with which the investigator is identified,
and the social-moral pressures of the society in which the investi-
gator lives. This last factor obviously plays a more extensive (and
more manifest) role in research in the social and behavioral

\textsuperscript{*} See the survey by Ort\textsuperscript{88} for a comprehensive review of the concept and the litera-
ture to that date.

\textsuperscript{†} Indeed, this is one way to conceptualize a major function of controls in research.
For fuller discussion of this issue, see GAP Report No. 42, CONTROLS IN PSYCHIATRIC
RESEARCH.\textsuperscript{89}
sciences, where the research is on things that matter most personally and intimately to people, than in other fields of research.‡

Most investigators will be swayed by the pressure of expectations of others, but to differing degrees, and with wide variations in regard to their sensitivity to different kinds of pressures. And each investigator may react differently at different times or in different situations. An investigator's excessive need to please will make him exquisitely sensitive to expectations and pressures and may have a major effect on his observations and measurements.§ While the need for finding the expected is almost universal among investigators, there are other important personality traits—occurring perhaps with lesser frequency—that can bring out an opposite subjectivity and bias in research work. Among them are an excessive need to find the unexpected, a fear of conformity, and an overweening self-enhancement.

Transcending the idiosyncratic motivations and countermotivations of individual scientists in influencing perception and apperception, and hence the possibility for new discovery, are the larger sweeps of the cumulative scientific enterprise itself. There are more open and more closed phases in the history of science; times of innovation and times of reliance on authority and of consolidation of tested knowledge. Nineteenth century physics represented one such relatively closed system. At the time, many thought that the physics of the Newtonian world view had reached such a state of perfected development that new knowledge accrued by investigation would play a progressively smaller and more peripheral role to the central core of already estab-

‡ Illuminating in this connection is the undoubted influence of the social, political, and moral position one enjoys on the strivings of Negro Americans for full equality of treatment: (civil rights, school desegregation, etc.) in relation to one's assessment of data on such issues as that of the existence of innate intellectual differences between the races or that of the effects of integration on students in bitherto segregated school systems or that of the effect on property values of selling houses in bitherto all-white neighborhoods to minority group members. This human equation is of course a different issue from the human equation involved in the different figures arrived at by two physicists both measuring the velocity of light (though this last is also a value-influenced issue).

§ One of the earmarks of the great originator in research is his ability not to be swayed by the expectations of others with regard to his work.
PROBLEMS OF DESIGN AND OF CONFIRMATION

Research Strategies

There are various approaches to research design; each of them has important implications for the kinds of problems that can be set, the kinds of questions that can be asked and answered, and the kinds of changes that can be assessed. A major distinction is between the longitudinal and the cross-sectional approach. These approaches have differing spheres of special applicability: the former to the study of intrasubject change (across time), the latter to the study of intersubject change (across space). Depending on the research problem, various combinations are possible, including additively the longitudinal following of a whole population, with assessment of each individual member cross-sectionally at specified points in time, permitting within the same design intrasubject and intersubject comparison. Less immediately obvious is another possible permutation that enables the researcher to approach a longitudinal study within a cross-sectional time compass by sampling across a wide population chosen as homogeneous in terms of its defining elements, but varying systematically in age. That is, within a brief time span one might be able to infer how individuals might change over a long time span in the characteristics that are the specific objects of research inquiry.

Clearly such choices are not possible in all circumstances. They are often useful (and almost essential) if one is to study definitively aspects of growth and development, or factors facilitating and/or inhibiting the evolution of a pathological process that may take place over more than the investigative life span of the researcher. Thus cross-sectional studies of large numbers of individuals of all age groups may provide vital information concerning the psychological vicissitudes associated with evolving disease processes—for example, psychosomatic illnesses.60 Or, cross-sectional assessment of the deployment of defensive and adaptive capacities at different age-levels will throw significant light on the various coping strategies that children employ to handle experiential vicissitude. The caution must of course be kept in mind that such cross-sectional inquiry as a way of studying longitudinal change runs the risk of blurring important aspects of longitudinal change typical of or peculiar to only certain individuals.

Another frequent distinction in research strategies is that between the idiographic and the nomothetic approaches, that is, between idiosyncratic personal and normative group phenomena.61-65 This is not merely a restatement of the distinction between the longitudinal and the cross-sectional approaches. Nomothetic studies can be longitudinal, following group norms and deviations over time, and an idiographic approach can be cross-sectional, as in the intensive study of an individual's psychological functioning at a given point in time.

Still another frequently made distinction is that between the naturalistic and the experimental approach. Here the discrimination is between the effort to observe the phenomena as they occur in all their "natural" complexity, as against the effort to isolate out aspects of the field for more controlled study by manipulations that hold some aspects constant while allowing others to vary systematically. This difference is easy to see as between the ethologist observing birds in their natural habitat and the experimentalist manipulating them in operant conditioning machines. But the distinction is by no means always so clear-cut. The subject of much dispute among researchers is the point at which the introduction of devices to enlarge the sphere of recorded and remembered data (sound recordings, one-way-vision watching, color movies, etc.) and of devices to tap additional data sources that do not ordinarily enter into the transaction (interviews with
the therapist to learn why he said what he did, what he thought but chose not to say and why, etc.) makes psychotherapy no longer a "natural" process, but a systematically (i.e., experimentally) altered process. In this situation, there always remains the question, "How much have the observed phenomena and the changes assessed in them been experimentally modified from their presumed natural state?"

Another somewhat related distinction is between clinical and statistical research approaches. But again, this is not the same as the difference between naturalistic and experimental. Both the naturalist and the experimentalist can be operating in the clinic, and both can be of a varying degree statistical in their approach, although the experimentalist is usually more concerned with questions of statistical control. But in some situations of naturalistic research, it is precisely the inability to carry out experimental manipulation that forces greater recourse to statistical control over inference.

A last distinction to be mentioned is that between "forward" and "backward" studies of change. Much valuable research has been done by way of retrospective study of available records, and this is the usual method of the clinical case report. Nonetheless, where it is possible—and especially in systematic formal research—forward study is to be preferred on a number of counts. Many types of retrospective anamnestic data are of very questionable reliability. It is generally difficult, if not impossible, to make maximal sense and use of previously collected data, even though these data may be reliable. And it is of course preferable to plan any research study in terms of relevant hypotheses, samples, variables, measures, and the like that are only possible when one begins with a planned study forward and then assesses change over an ensuing period. Hypotheses that do emerge from backward study of data are less "virile" than hypotheses that were incorpo-

† For a presentation of the varying viewpoints on this issue of the place of prediction (and postdiction, and also "co-diction") in relation to clinical research in psychiatric treatment and in psychoanalysis, see the discussion by Wallenstein in his article on the role of prediction in theory-building in psychoanalysis.
replication allows a statistical estimate of error and thus guarantees the validity of the tests of significance.\textsuperscript{105}

Replication has been defined with a somewhat different emphasis by Winer:\textsuperscript{106}

A replication of an experiment is an independent repetition under as nearly identical conditions as the nature of the experimental material will permit. \textit{Independent} in this context implies that the experimental units in the repetition are independent samples from the population being studied. That is, if the elements are people, in a replication of an experiment an independent sample of people is used in the replication.

In this definition, replication is distinguished from repetition; in replication independent samples are employed, while in repetition the same samples are used again under all the treatment conditions.

As an example of an experiment involving both repeated measurements and replication, we may consider the situation in which we wish to measure the effects of four drugs upon five subjects. The following design would meet these conditions:

<table>
<thead>
<tr>
<th>Subject</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>xA1</td>
<td>xB1</td>
<td>xC1</td>
<td>xD1</td>
</tr>
<tr>
<td>2</td>
<td>xA2</td>
<td>xB2</td>
<td>xC2</td>
<td>xD2</td>
</tr>
<tr>
<td>3</td>
<td>xA3</td>
<td>xB3</td>
<td>xC3</td>
<td>xD3</td>
</tr>
<tr>
<td>4</td>
<td>xA4</td>
<td>xB4</td>
<td>xC4</td>
<td>xD4</td>
</tr>
<tr>
<td>5</td>
<td>xA5</td>
<td>xB5</td>
<td>xC5</td>
<td>xD5</td>
</tr>
</tbody>
</table>

Going horizontally, the testing of the four drugs on any single subject involves repeated measurement of the same subject under all the treatment conditions. Vertically, the testing of any single drug on each of the five subjects involves replication of the single treatment. If, having performed the experiment with the five subjects and the four drugs, the investigator chooses five additional subjects and repeats the entire experiment using the same drugs, the entire experiment will be said to have been replicated.

The distinction between an experiment involving repeated measurements and one involving replication is not always apparent and may depend upon the purpose of the experiment and upon the hypothesis to be tested. For example, if a given subject is weighed on five different scales in order to obtain a best estimate of his true weight, the experiment involves repeated measurement. If he is weighed five times on the same scale for the same purpose, it would also be a case of repeated measurement. However, if the purpose of the latter experiment is to obtain an estimate of the variability in the scales, each weighing may be considered to be a case of replication. The statistical treatment of designs involving repeated measurement differs from those involving replications.\textsuperscript{106}

The purpose of repeated measurement and of replication is very similar. Both procedures attempt to increase the reliability and generalizability of experimental data by utilizing repetitive techniques. Both are used to separate subject variability from treatment effect and from measurement error. The repeated measurement experiment using the same subject does this by requiring that each subject serve as his own control. Responses of individual subjects to the treatment are measured in terms of deviations about a value that marks the average response of that individual subject. Thus, one source of variability (in this case, intra-subject variability) is eliminated from the experimental error. Replication experiments attempt to reduce subject variability by using a sufficiently large number of independent subjects studied repetitively under as nearly identical conditions as possible. The assumption underlying this method is that the selected subjects will include representative samples of the population.

The confidence associated with inferences based upon a series of replications is always greater than the confidence associated with the inferences made from any of the single experiments constituting the series. Furthermore, in replicating experiments, conditions are never fully identical. For example, time and/or place presumably vary. Replication will therefore demonstrate whether the effects of extraneous variables are insignificant as compared
to treatment effects; greater generalization of the initial findings may consequently be made. That is, even if replication is unsuccessful, something is salvaged and knowledge accrues.

These “hard” definitions of replication derived from modern statistics have been broadened to include a host of repetitive techniques whose common purpose is to increase the reliability and generalizability of experimental data. Zetterberg, a sociologist, proposes such a broad view of replication:

If a single term should be singled out for sociological verification, it is perhaps replication. Replication can be performed in many senses and we would like to recommend them all. The one most overlooked is replication in the sense of testing different implications of the same theory. A consistency found between such tests may give the theory quite high plausibility. Furthermore, we may encourage replications using different operational definitions. This should reveal something about the reliability of our procedures and the validity of our constructs. Also we may replicate in different populations to insure the universality of our theory. Furthermore, we can perform replications in the statistical sense, adding more cases to our data to be sure we have refuted the null hypothesis. Finally when using experimental designs, we replicate tests controlling different factors in different replications, thus minimizing the probability that alternative hypotheses account for the findings we claim as support for the hypotheses of our theory.

Broadened to this extent, the concept of replication may seem to encompass almost all aspects of inductive and deductive reasoning.

Sidman has classified the forms of replication most useful in psychological research. He suggests the term “direct replication” for those techniques that repeat an experiment with new subjects or make repeated observations on the same subjects under each of several experimental conditions. “Systematic replication,” on the other hand, uses data gathered under one set of specified conditions as the basis for performing new experiments under new conditions to obtain additional related data. Thus the investigator purposely varies conditions to test whether a relationship occurs under conditions different from those of the original experiment.

Whether or not replication is employed in a narrow or broad definition, it is clear that it is a major feature of experimental methods. Direct replication offers an empirical test of the reliability of observations. Systematic replication permits the testing of the generality of a proposition or theory. Replication with different operational definitions reveals much about the validity of constructs. Replication with systematic variation also has heuristic value, permitting the separation of essential conditions from irrelevant ones and promoting the generation of new information and new theory. If the creation of general propositions in the form of facts, theories, and laws is the aim of scientific endeavor, replication in the broad sense offers a powerful technique for achieving this aim.

**Direct Replication**

By this technique the investigator attempts to duplicate any single experiment as closely as possible. This is largely a test of reliability. Ignoring, for the moment, replication by different investigators or with different instruments, direct replication may be intersubject (or intergroup) or intrasubject (or intragroup). When an experiment has been performed once with a single subject, the investigator may choose to add additional subjects, to repeat the experiment at a different time with the same subject, or do both. If he chooses intersubject replication by increasing his population, he is attempting to determine whether his variables are adequately controlled or whether uncontrolled or unknown variables might prevent successful repetition. If he employs the technique of intersubject replication, his work will still be subject to the scrutiny of other investigators, who will attempt to replicate it (interinvestigator replication) under closely similar conditions, and will also be subject to scrutiny for estimates of internal consistency under specifically varied conditions, by Systematic Replication. If the findings are confirmed by investigators
of different theoretical persuasions, their credibility is even further enhanced.

Sidman prefers to use intersubject replication rather than intergroup replication. In the latter, a group of subjects is used as an n of 1. Intergroup replication provides an indicator of reliability only insofar as it shows that changes in central tendency can be repeated. But it is less useful for generalization because it does not tell how many individuals the data actually represent. The use of populations for intergroup replication may also create a conceptual hazard because significant findings relevant to subgroups may be canceled out when the total group is considered as a unit. A clear illustration of this may be found in the area of mental retardation. Blood levels of phenylalanine in a large group of retardates lie within normal limits, and studies treating retardates as a whole would fail to note the subpopulation of phenylketonurics, whose blood phenylalanine level is significantly elevated. Similar considerations may apply to the biology of schizophrenia; and one cannot help but wonder, in regard to schizophrenia research, where biological data have for years been demonstrated to lie within the normal range, whether there may not be some fault in the sense of obscuring meaningful subgroups of schizophrenic patients through considering them all as a whole.

If the investigator chooses intrasubject or intragroup replication—which, as already indicated, is considered by some statisticians to be more properly viewed as repeated measurements—the technique is that of repeated manipulation of an individual subject or group in a consistent fashion. When the measure remains consistent, the investigator may be certain that the relevant variables are well under control. Requirements for intrasubject replication are that there be a definable baseline, that the changes induced by the experiment be reversible, and that there be no significant interaction effect (no significant learning) from repeated exposures. A defined baseline need not be constant, provided that it changes in an orderly and replicable fashion, but such changes must be independent of the particular operation that is superimposed on the baseline; if they are not, then intrasubject replication will not be possible.

When intrasubject replication is possible, the investigator has a powerful technique for demonstrating the reliability of his observations. Intersubject variability is automatically eliminated, smaller effects may be reliably observed than it is possible to observe with group techniques, and new phenomena that might be masked by the use of groups may be uncovered. Thus, if one were to study the phenylalanine blood levels in a representative cross-section of mental retardates, the mean value for this group would not differ significantly from that of the population at large. But if one were to focus on the particular individual whose phenylalanine is elevated and find that it is consistently elevated in a series of intrasubject replications, one has located a new disease entity.

A problem occasionally arises where intrasubject replication is not possible but intersubject replication is. This will occur when original baseline values cannot be reestablished because the experimentally induced changes are not reversible, in which event recourse must be taken to intersubject replication. Learning experiments are a good example of this situation.

A knotty question may be posed in deciding upon the number of individuals upon whom an intersubject experiment must be replicated before the data may be considered representative. Equally, in intrasubject replication one may ask how many times an experiment must be replicated on a single subject to obtain a given degree of confidence in the reliability of one's findings. Sidman points out that a complete "statistics of replication" is not yet available and that many factors must be considered in trying to find an answer to such a question. Thus, if new data disagree with old, well-established data, more replications will be required than might otherwise be the case. The ability of the investigator is important; experimental data are rarely independent of the experimenter, and an investigator with well-established

---

\[\text{\footnotesize \textsuperscript{1}}\text{In intrasubject learning experiments, you are repeating but are not replicating.} \]
skills and techniques may need or decide to perform fewer replications. The variability of the data influences the number of required replications. So also does the confidence in the experimental control and the magnitude of the experimental effects.

Additionally, since all instances are not the same, this is not purely a statistical problem. Also operative are questions of social value and of personal idiosyncrasy. For example, the importance of the research and the consequences if the findings are correct or incorrect will be significant determinants of the number of desirable replications. If one is seeking to establish a constant or a baseline upon which much other research will depend and for which very high accuracy is required, the number of replications required will be greater than that required for the establishment of an isolated experimental datum. Similarly, personal attributes, like the authority and fame of the investigator or, on the other hand, how self-critical he is, can decrease or increase the number of replications sought.

**Systematic Replication**

Systematic replication as used by Sidman may also be considered among the general class of techniques employed by investigators for hypothesis testing. When an investigator has obtained a series of observations and finds an orderly relationship between the operations of the experiment and the observed phenomena, he will wish to estimate both the reliability of the findings and their generalizability as well. The reliability may be estimated either by direct replication or by systematic replication. The method chosen is a judgment of the investigator. There are no well-specified rules to follow, nor is there any satisfactory classification of the types of systematic replication. Usually the choice to replicate systematically is based upon the nature of the investigator’s curiosity and imagination. If he has confidence in his techniques and wishes to extend the generalizability of his findings while at the same time increasing his reliability, he will choose systematic replication. Instead of merely repeating his experiment he will use the data as a basis for performing new experiments to obtain additional related data. In a systematic replication, the investigator may deliberately refrain from controlling certain variables and if the procedure is successful, he has found these variables to be irrelevant because the original finding is sustained under conditions different from those of the original experiment.

There is often a gamble involved in the decision to perform a particular systematic replication rather than a direct replication as the next research step. If one is successful, time, effort, and money are saved and generalizability is greatly extended. If one fails, the original experiment will have to be redone because there is now no way of knowing whether the failure stems from inadequate control of relevant variables or whether new variables were introduced in the systematic replicating experiment. The gamble is minimized through the fullest utilization of the already existing knowledge in the field. This knowledge may consist of the investigator’s acquaintance with the fundamental laws of his science, his detailed acquaintance with his own previous work and that of fellow investigators, or merely a knowledge in depth of his instruments and of his repertoire of skills in using them. The results obtained in systematic replication must also be exposed to the existing body of related knowledge to determine their fit.

Sidman describes three techniques of systematic replication: (1) the baseline technique, (2) the probe technique, and (3) the “affirming of the consequent” technique. The baseline technique is closest to direct replication insofar as a given phenomenon is used as a baseline for the investigation of other variables. Thus, if in the course of research a phenomenon (e.g., a type of behavior) is generated and found to be reliable, it may in itself be used to investigate other problems. For example, if an animal is conditioned to respond in a given fashion and a procedure is found that disrupts this response, the disruption may be used to investigate a whole series of new procedures that may or may not also disrupt the response. In testing the new procedures, the old findings will be replicated over and over. Manifestly the baseline
technique requires a definable base, and preferably this base should be at an intermediate range so that shifts up or down from the original value may be obtained.

The probe technique of systematic replication may be employed to seek out unifying variables that are common to a set of related experiments. For this technique to be used, there must be a definable baseline and a phenomenon that the investigator can produce at will and as often as he desires. Experimental conditions are then varied in an expanding fashion, not to create final experiments but rather to increase the scope of information about the phenomenon under observation. With this procedure the investigator seeks to define the limits of generalizability of the underlying hypothesis and simultaneously to assess the reliability of the findings.

Replication by affirming the consequent represents another short cut. Given the proposition that if A is true, B will be true (where A is a sufficient but not necessary condition for B), we experimentally determine that B is true. From this we might conclude that A is true. Since the proposition did not state that B is true only if A is true, A will not necessarily be true even when B is found true. However, this technique may be used inductively by the investigator because it does tell us something about A. Thus if B is found to be false, A must necessarily also be false. Since B has been found to be true, the likelihood that A is false diminishes somewhat. If the experiment is repeated with propositions C, D, E, and so forth, the plausibility that A is correct will constantly increase. All prediction in personality theory, in psychiatric treatment, and in psychotherapy is, in fact, based on this model, since if predictions are correct, they are not necessarily correct for the reasons adduced, but if they are incorrect, then clearly something is awry in the chain of evidence and assumptions that underly and form the basis for the prediction. Of this, Wallerstein\textsuperscript{19} says,

Thus the outcome of no single prediction, nor even of any given series of predictions underpinned by the same theoretical proposition, will provide a crucial and definitive test of the adequacy of

that proposition; but certainly predictive success, and repeatable success, will strengthen the credibility of that proposition, and predictive failure will weaken it.

In the absence of logical proof, however, this method of affirming the consequent will not persuade adherents of alternative explanations or theories of the correctness of the evidence. Nonetheless, as the plausibility of correctness increases, the systematic structure that has been created provides a basis for further experimentation and systematization. So long as the newer findings remain internally consistent and have pragmatic value, they will become incorporated into the body of accepted knowledge.

Systematic replication has great utility when dealing with irreversible changes that prevent direct replication with a single organism. Instead of attempting to replicate the original findings, the implications of the data are studied. Systematic replication is also useful when dealing with historical techniques, as in paleontology and psychoanalysis. Here the original conditions for the creation of a phenomenon can rarely be replicated, but the implied consequences can be scrutinized and tested. Finally, the method is used in theory testing, for if the consequences are negative, the theory must be modified; if positive, its plausibility is increased.

In conclusion, it is worth emphasizing that replication in its purest sense is never totally possible, since, if time is considered, nothing in the universe repeats itself identically. Nothing has the same age at two different times, and no living thing is composed of the same molecules at two different times. And if two operations are carried out simultaneously in time, then there is a difference in locus or space. Replication is therefore an ideal that operationally we only approximate.

Since total replication is an ideal state, we may choose to come as close as possible to it, as in direct replication, or to deviate from it deliberately in specified directions, as in systematic replication. If we choose the former tactic, we immediately, even if unwittingly, make judgments as to relevant and irrelevant vari-
ables. Such judgments are based on previous concepts and previous information, and mold significantly the design of the replication experiment. Thus, if we measure the speed of light or the performance of a human subject on an IQ test and then wish to repeat it, we are not likely to consider the day of the week a significant variable. If we find different results on different days, we would on the basis of previous experience likely judge that the discrepancy is due to instrumental error rather than to alterations in the phenomenon under investigation. On the other hand, if we were to count the number of cars passing a given spot and then on repeating the count a few days later find a different value, we would on the basis of previous experience be more likely to consider the difference as due to the day of the week or the hour of the day rather than to the counting instrument.

In either case, the data derived from a replication of the experiment reflect upon the hypothesis we entertain and permit the specification of conditions relevant to it, thus further refining the hypothesis. This is achieved by a random sampling of replications in which the irrelevant variables have been gradually eliminated. Thus, if the rate of fall of a body is independent of its composition, color, shape, smell, size, and weight and independent of the temperature, humidity, wind direction, or day of the week, we can generalize as Galileo did. On the other hand, if in a series of replications certain factors are found to significantly alter the findings, the relevance of these factors to the hypothesis in question is established.

Cohen and Nagel describe the purposes of repeated experiments in terms of the circular method of science as follows:

1. In virtue of its method, the enterprise of science is a self-corrective process. It appeals to no special revelation or authority whose deliverances are indubitable and final. It claims no infallibility, but relies upon the methods of developing and testing hypotheses for assured conclusions. The canons of inquiry are themselves discovered in the process of reflection, and may themselves become modified in the course of study.

2. General propositions can be established only by the method of repeated sampling. Consequently, the propositions which a science puts forward for study are either confirmed in all possible experiments or modified in accordance with the evidence. It is this self-corrective nature of the method which allows us to challenge any proposition, but which also assures us that the theories which science accepts are more probable than any alternative theories. By not claiming more certainty than the evidence warrants, scientific method succeeds in obtaining more logical certainty than any other method yet devised.

3. In the process of gathering and weighing evidence there is a continuous appeal from facts to theories or principles, and from principles to facts. For there is nothing intrinsically indubitable, there are no absolutely first principles, in the sense of principles which are self-evident or which must be known prior to everything else.

4. The method of science is thus essentially circular. We obtain evidence for principles by appealing to empirical material, to what is alleged to be “fact”; and we select, analyze, and interpret empirical material on the basis of principles. In virtue of such give and take between facts and principles, everything that is dubitable falls under careful scrutiny at one time or another.

The Use of Controls

Control, testing, or limiting the various sources of error and of distortion is a crucial component of the entire scientific enterprise and of course operates in every aspect of the assessment of change. The previous GAP report, CONTROLS IN PSYCHIATRIC RESEARCH, prepared by this committee was devoted to an explication of the relationship of the concept of control to the broad principles that govern the scientific process and to the implementation of this concept in specific psychiatric research techniques and procedures.

In the assessment of change, the investigator has at his dispos-
al several types of controls, which can of course be admixed. For the purposes of this discussion, these may be subsumed under the headings—experimental, sampling, and statistical. In using experimental controls, the experimenter imposes conditions that change, manipulate, or otherwise influence the phenomena he is studying. This is the classical experimental treatment approach. Another type of control is, in a way, intermediary between the experimental and the statistical: it involves the use of proper selection, that is, sampling. For example, the investigator may select two different groups of subjects that represent differing instances, or differing conditions, of the phenomena under study. That is, rather than direct experimental manipulation by the investigator, instances are selected of existing differences, "experiments in nature," study of which highlights the problem at issue. In statistical controls, the investigator does not try to manipulate the phenomena at all, not even via the search for particular instances or particular conditions; rather, he measures existing phenomena and by statistical means removes, modifies, or otherwise accounts for their effects.

An example of an alternate sampling or statistical control model in relation to an issue discussed at length earlier herein (see page 414: "Problems of Measurement: The Phenomenon as the Variable") is the following: in an experiment in which initial level of scores (or baseline) is an important determinant of the change score, the investigator might run separate analyses for all persons having comparable initial scores (sampling control) or, by appropriate statistical techniques, partial out the effect of the initial score on the change score (statistical control). For expanded discussion of this and related issues in regard to both psychosocial controls and formal, technical controls, the reader is referred to the GAP report mentioned above.⁶⁰

The proper assessment of change is a core problem of clinical research in any aspect of psychiatric science—as indeed of all science. Its very ubiquity and commonplace nature as a research task mask, we feel, the very complex issues involved in such assessment, issues that we have tried to clarify throughout the entire discussion in this report. These many issues have been presented under a variety of headings: (1) The determinants of research choices about what to assess and how to measure, in terms of personal value orientations and commitments; sociocultural influences; the state of development of the science; the prevailing scientific models and identifications; and the different philosophies inherent in the varieties of scientific research values. (2) The meanings of the concept of change, in terms of its contextual qualities, its form varieties, and its dimensions. (3) The various factors that influence the assessment of change. (4) The many problems of theory and practice in measurement (as a major means of assessment) again from a variety of points of view—the phenomenon, the setting, the subject, the investigator, and the instrument, each considered as the variable to be measured. (5) Problems of design and of confirmation, such as the available research strategies, methods of replication, and methods of control.

In regard to each of these many complex issues, we have expounded two basic themes. The first is that in the problems of psychiatric research the basic clinical truism of our psychological science is equally compelling—that things are not necessarily to be taken at face value, that they may be other than what they seem, more or less in quantity, simpler or more complex in quali-
ty. A statement of a difference (i.e., of a change) is not necessarily either a precise or a valid rendering of a true difference in the phenomenon under study. The second basic—and corollary—theme is that many judgments, assessments, and measurements purporting to state changes could be rendered more precise and more valid by attention to the variety of factors discussed in this report and that psychiatric research could thereby profit.

As a last point, all of these considerations and our two guiding themes are equally applicable to each one of the whole sequence of steps in the research enterprise; from the first theoretical hunches, through the formulation of research questions and hypotheses, the selection of variables and of instruments, the choosing of measurement techniques, the sampling of subjects, the total design of the experiment and then the methods for its statistical analysis, the interpretation of the research findings, and the final relating of the results back to the initial starting point in theory, in the form of new research questions and hypotheses.

This total range of requirements may be beyond the capabilities of any single investigator. This does not mean, however, that the importance of carefully considering each requirement, and of the expertise brought to bear on it, can be minimized or ignored. Being human, an investigator may be inclined to underestimate the importance of, and the complexities involved in, aspects of the research process with which he may not be intimately experienced and knowledgeable. A particular investigator may need to seek outside help in order to be certain that all the links in the chain of research steps are strong, and, so far as possible, equally strong. Too often one link is strong, but at the expense of the others; or the strength of one link can mislead one into thinking that all the links must somehow be strong. As a result, an investigator may base his conclusions on overly simple or otherwise inappropriate procedures and so unwittingly reach conclusions that are unnecessarily fallible—that contain an unnecessarily large component of error. If this report serves to diminish this error component in research in our field, it will have served its purpose.

REFERENCES

53. Rogers, Carl; and Dymond, Rosalind F.; Psychotherapy and Personality Change, University of Chicago Press, Chicago, 1954.
80. Gardner, Riley W.; Holzman, Philip S.; Klein, George S.; Liofaff.


Permissions

The members of the GAP Committee on Research appreciate the permission to quote in this report from the following sources:


Acknowledgments

The program of the Group for the Advancement of Psychiatry, a non-profit tax exempt organization, is made possible largely through the voluntary contributions and efforts of its members. For their financial assistance during the past fiscal year, in helping it to fulfill its aims, GAP is grateful to the following foundations and organizations:

Sponsors

Bing Fund
Ciba Corporation
Maurice Falk Medical Fund
Geigy Chemical Corporation
General Service Foundation
The Grant Foundation
The Hoffmann-La Roche Foundation
Parker Davis & Company
Smith Kline & French Foundation
Taconic Foundation
Upjohn Company
Wallace Laboratories
Wyeth Laboratories

Donors

Carrier Clinic
The Louis and Pauline Cowan Foundation, Inc.
Forrest Hospital Foundation
Gralnick Foundation
Morris and Sophie Kardon Foundation
Riker Laboratories
Publications of the
Group for the Advancement of Psychiatry

Because readers of this publication may not be aware of previously published GAP Reports and Symposia, a selected listing of titles is given below.

<table>
<thead>
<tr>
<th>Number</th>
<th>Title</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>S-2</td>
<td>*Illustrative Strategies for Research on Psychopathology in Mental Health—July 1956</td>
<td>$ .40</td>
</tr>
<tr>
<td>S-3</td>
<td>*Factors Used To Increase the Susceptibility of Individuals to Forceful Indoctrination: Observations and Experiments—Dec. 1956</td>
<td>.40</td>
</tr>
<tr>
<td>S-4</td>
<td>*Methods of Forceful Indoctrination: Observations and Interviews—July 1957</td>
<td>.75</td>
</tr>
<tr>
<td>38</td>
<td>The Diagnostic Process in Child Psychiatry—Aug. 1957</td>
<td>.50</td>
</tr>
<tr>
<td>42</td>
<td>Some Observations on Controls in Psychiatric Research—May 1959</td>
<td>1.00</td>
</tr>
<tr>
<td>50</td>
<td>Problems of Estimating Changes in Frequency of Mental Disorders—Aug. 1961</td>
<td>.75</td>
</tr>
<tr>
<td>S-8</td>
<td>*Medical Uses of Hypnosis—April 1962</td>
<td>1.00</td>
</tr>
<tr>
<td>52</td>
<td>The College Experience: A Focus for Psychiatric Research—May 1962</td>
<td>1.00</td>
</tr>
<tr>
<td>S-9</td>
<td>*Pavlovian Conditioning and American Psychiatry—March 1964</td>
<td>1.00</td>
</tr>
<tr>
<td>57</td>
<td>Psychiatric Aspects of the Prevention of Nuclear War—Sept. 1964</td>
<td>1.50</td>
</tr>
<tr>
<td>62</td>
<td>Psychopathological Disorders in Childhood: Theoretical Considerations and a Proposed Classification—June 1966</td>
<td>3.50</td>
</tr>
</tbody>
</table>

* "S" refers to Symposium

A complete listing of publications of the Group for the Advancement of Psychiatry may be obtained upon request from the Publications Office.

Bound volumes of Reports and Symposia published since 1947 are also available. They include reports that are now out of print and unavailable in any other form.

This report was produced for the Group for the Advancement of Psychiatry by the Mental Health Materials Center, Inc., New York.