COMPREHENSIVE HANDBOOK OF PERSONALITY AND PSYCHOPATHOLOGY

VOLUME 2
ADULT PSYCHOPATHOLOGY

Frank Andrasik

Volume Editor

Michel Hersen

Jay C. Thomas

Editors-in-Chief

John Wiley & Sons, Inc.

COMPREHENSIVE HANDBOOK OF PERSONALITY AND PSYCHOPATHOLOGY

VOLUME 2
ADULT PSYCHOPATHOLOGY

Frank Andrasik

Volume Editor

Michel Hersen

Jay C. Thomas

Editors-in-Chief

John Wiley & Sons, Inc.

COMPREHENSIVE HANDBOOK OF PERSONALITY AND PSYCHOPATHOLOGY

VOLUME 2
ADULT PSYCHOPATHOLOGY

Frank Andrasik

Volume Editor

Michel Hersen

Jay C. Thomas

Editors-in-Chief

John Wiley & Sons, Inc.

This book is printed on acid-free paper. ♾

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:

Comprehensive handbook of personality and psychopathology / Michel Hersen & Jay C. Thomas, editors-in-chief.
p. ; cm.
Includes bibliographical references.
ISBN-13 978-0-471-47945-1 (cloth : alk. paper : set)
ISBN-10 0-471-47945-4 (cloth : alk. paper : set) —
ISBN-13 978-0-471-48837-8 (cloth : alk. paper : v. 1)
ISBN-10 0-471-48837-2 (cloth : alk. paper : v. 1) —
ISBN-13 978-0-471-48838-5 (cloth : alk. paper : v. 2)
ISBN-10 0-471-48838-0 (cloth : alk. paper : v. 2) —
ISBN-13 978-0-471-48839-2 (cloth : alk. paper : v. 3)
ISBN-10 0-471-48839-9 (cloth : alk. paper : v. 3)
1. Psychology, Pathological—Handbooks, manuals, etc. 2. Child psychopathology—Handbooks, manuals, etc. 3. Personality—Handbooks, manuals, etc. 4. Psychology—Handbooks, manuals, etc. I. Hersen, Michel. II. Thomas, Jay C., 1951–
[DNLM: 1. Mental Disorders—therapy. 2. Personality. 3. Psychological Theory. WM 400 C737 2006] 1951–
RC456.C66 2006
618.92′89—dc22

2005043981

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

Contents

PART THREE
TREATMENT APPROACHES

Handbook Preface

Remarkably, the linkage between personality and psychopathology, although extensive, has not been underscored in the larger tomes on these subjects. In the last decade there have been many books on personality, adult psychopathology, and child psychopathology, but none seems to have related the three in an integrated fashion. In part, this three-volume *Comprehensive Handbook of Personality and Psychopathology* (CHOPP), with the first volume on *Personality and Everyday Functioning,* the second on *Adult Psychopathology,* and the third on *Child Psychopathology,* is devoted to remedying this gap in the literature. Another unique feature of CHOPP appears in the volumes on *Adult Psychopathology* and *Child Psychopathology,* where impact of adult and child psychopathology on family, work, school, and peers is highlighted, in addition to the relation of specific psychopathology to normal development. Given the marked importance of such impact, contributors were asked to delineate the negative impact of psychopathology on the individual's daily environments.

In light of the aforementioned features, we trust that CHOPP is timely and that it will be well received in many quarters in psychology. The work should stand as an entity as a three-volume endeavor. However, given the structure of each volume, we believe that it is possible to break up the set into individual volumes for relevant courses on personality, normal development, adult psychopathology, and child psychopathology.

Volume 1 (*Personality and Everyday Functioning*) contains 23 chapters divided into four parts (Foundations, Broad-Range Theories and Systems, Mid-Range Theories, and Special Applications). This volume is unique in that it encompasses both the broad theories of personality and those theories with a more limited range, known as mid-range theories. Broad-range theories were originally developed to explain the behavior of normal people in everyday situations. But it also is important to have a reference point for those individuals suffering from various sorts of psychopathology. Chapters in this section follow a general format where possible:

A. Statement of the Theory
B. Developmental Considerations
C. Biological/Physiological Relationships
D. Boundaries of the Theory
E. Evidence in Support of and against the Theory
F. Predictions for Everyday Functioning
    1. Family Life
    2. Work or School
    3. Retirement
    4. Recreation

Thus, Volume 1 sets the stage for Volumes 2 and 3 while at the same time standing on its own for understanding everyday life from the personality perspective.

Volume 2 (*Adult Psychopathology*) contains 30 chapters divided into three parts (General Issues, Major Disorders and Problems, Treatment Approaches). Volume 3 (*Child Psychopathology*) contains 27 chapters divided into three parts (General Issues, Major Disorders and Problems, Treatment Approaches). As previously noted, a unique feature in these volumes is mention of the impact of psychopathology on the family, work, school, and peers, often neglected in standard works. In both Volumes 2 and 3, most of the contributors have adhered to a relatively standard format for Part Two. In some instances, some of the authors have opted to combine sections.

A. Description of the Disorder
B. Epidemiology
C. Clinical Picture
D. Etiology
E. Course, Complications, and Prognosis
F. Assessment and Diagnosis
G. Impact on the Environment
    1. Family
    2. Work or School
    3. Peer Interactions
H. Treatment Implications

In addition, authors in Volume 3 include the sections Personality Development and Psychopathology and Implications for Future Personality Development. We trust that the relatively uniform format in Part Two of Volumes 2 and 3 will make for ease of reading and some interchapter comparisons within and across volumes.

Many individuals have worked very hard to bring this series of volumes to fruition. First, we thank our editor at John

Wiley, Tracey Belmont, for once again understanding the import and scope of the project and having confidence in our ability to execute in spite of interfering hurricanes, other natural events, and varied life events. Second, we thank our editors of the specific volumes for planning, recruiting, and editing. Third, we thank our eminent contributors for taking time out from their busy schedules to add yet one more writing task in sharing their expertise. Claire Huismann, our project manager at Apex Publishing, deserves special recognition for her extraordinary efforts, competence, and patience throughout the creation of this series. And finally, but hardly least of all, we thank all at John Wiley and Pacific University, including Carole Londeree, Linda James, Alison Brodhagen, Greg May, and Cynthia Polance, for their excellent technical assistance.

Michel Hersen and Jay C. Thomas
Forest Grove and Portland, Oregon

Preface to Volume 2

Volume 2 continues the themes articulated in the first volume of this series. It contains 30 chapters, divided into three parts. Part One includes chapters that discuss diagnosis and classification and pertinent research issues as well as separate chapters that discuss the behavioral, cognitive, genetic, sociocultural, and biological factors that influence development.

Part Two includes 21 chapters that cover a broad spectrum of disorders, including anxiety, mood, schizophrenia, organic, personality, substance use, eating, psychophysiological, sexual dysfunction and deviation, and marital dysfunction. These chapters continue the focus on the linkage of personality and psychopathology and how this impacts the individual's social unit (family and peers) and performance in work, school, and leisure settings. Although authors were asked to give equal weight to all of these specific impacts on the environment, the available literature demanded varied coverage, with authors left at times only to point out deficits in our current knowledge and future avenues for research. These chapters additionally provide descriptions of the disorders and the clinical picture; review epidemiology and etiological theories; discuss the typical course, complications, and prognosis; outline the approach to assessment and diagnosis; and review the literature bearing on treatment and the attendant implications.

Part Three includes three individual chapters, each focusing more in depth on the most current general treatment approaches for the conditions reviewed—psychodynamic, cognitive behavioral, and pharmacological.

A volume of this scope and size could not be possible without the eminent scholars who gave so generously of their time, in the face of multiple competing demands, to draft the copy you see here. It is equally true that a number of people worked just as diligently, behind the scenes, in order to produce this volume. My first word of thanks to the "behind-the-scenes crew" goes to the series editors, Michel Hersen and Jay C. Thomas, for affording me, and having the confidence in me, to serve as the volume editor and for providing assistance beyond that normally needed when Ivan the Terrible raised its ugly head. I thank Gayle Beck and Tim Brown for their wise consultation as I was selecting authors and topics. Prior to working on this volume, I felt I had a good handle on the editing process; however, working closely with Michel taught me that I had much to learn. I offer him my further thanks for teaching me so much more about the intricacies of successful editing. Final words of thanks are owed to Tracey Belmont and Isabel Pratt, both of John Wiley & Sons, for their patience, understanding, support, and flexibility; and to Claire Huismann, project manager at Apex Publishing, for her invaluable assistance, superb skills, and unflappable demeanor, all of which proved critical in getting this volume to the finish line in polished condition.

Frank Andrasik
Pensacola, Florida

Contributors

Frank Andrasik, PhD
University of West Florida
Pensacola, Florida

Martin M. Antony, PhD, ABPP
McMaster University
Hamilton, Ontario

Steven R. H. Beach, PhD
University of Georgia
Athens, Georgia

Alex Blaszczynski, PhD
University of Sydney
Sydney, Australia

Elisa Bolton, PhD
NH-Dartmouth Psychiatric Research Center
Concord, New Hampshire

Timothy A. Brown, PhD
Boston University
Boston, Massachusetts

Richard A. Bryant, PhD
University of New South Wales
Sydney, Australia

Laura Campbell-Sills, PhD
University of California, San Diego
La Jolla, California

Dorothy Chin, PhD
University of California, Los Angeles
Los Angeles, California

Meredith E. Coles, PhD
Binghamton University (SUNY)
Binghamton, New York

Eric W. Corty, PhD
Penn State Erie, The Behrend College
Erie, Pennsylvania

Timothey C. Denko, MD
University of Pittsburgh Medical Center
Pittsburgh, Pennsylvania

Keith S. Dobson, PhD
University of Calgary
Calgary, Alberta

Brad Donohue, PhD
University of Nevada, Las Vegas
Las Vegas, Nevada

Alisha M. Farley, BA
University of Nevada, Las Vegas
Las Vegas, Nevada

Frank Fincham, PhD
Florida State University
Tallahassee, Florida

Edna B. Foa, PhD
University of Pennsylvania School of Medicine
Philadelphia, Pennsylvania

Samantha L. French, BS
University of Nevada, Las Vegas
Las Vegas, Nevada

Drew Gouvier, PhD
Louisiana State University
Baton Rouge, Louisiana

Dolores V. Hernandez, MS
University of Texas at El Paso
El Paso, Texas

Marilyn Holmes, BA
Pennsylvania State University
University Park, Pennsylvania

Betty Horng, MA
Binghamton University (SUNY)
Binghamton, New York

Steven K. Huprich, PhD
Eastern Michigan University
Ypsilanti, Michigan

Kerry L. Jang, PhD
University of British Columbia
Vancouver, British Columbia

Charles Kamen, BS
University of Georgia
Athens, Georgia

Velma A. Kameoka, PhD
University of Hawaii at Manoa
Honolulu, Hawaii

Jennifer Karmely, MA
University of Nevada, Las Vegas
Las Vegas, Nevada

Rachel A. Keaschuk, MS
SUNY Upstate Medical University
Syracuse, New York

James Langenbucher, PhD
Rutgers, the State University of New Jersey
Piscataway, New Jersey

Kenneth N. Levy, PhD
Pennsylvania State University
University Park, Pennsylvania

Kristine L. Lokken, PhD
University of Alabama-Birmingham
Birmingham, Alabama

Elizabeth R. Lombardo, PhD, PT
Private Practice
Dallas, Texas

Erin Martinez, BA
University of Texas at Austin
Austin, Texas

Randi E. McCabe, PhD
McMaster University
Hamilton, Ontario

Susan R. McGurk, PhD
Dartmouth Medical School
Concord, New Hampshire

Kim T. Mueser, PhD
Dartmouth Medical School
Concord, New Hampshire

William D. Murphy, PhD
University of Tennessee Health Science Center
Memphis, Tennessee

Peter E. Nathan, PhD
University of Iowa
Iowa City, Iowa

Cory F. Newman, PhD
University of Pennsylvania, School of Medicine
Philadelphia, Pennsylvania

Michelle G. Newman, PhD
Pennsylvania State University
University Park, Pennsylvania

Arthur M. Nezu, PhD, ABPP
Drexel University
Philadelphia, Pennsylvania

Christine Maguth Nezu, PhD, ABPP
Drexel University
Philadelphia, Pennsylvania

Lia Nower, JD, PhD
University of Missouri-St. Louis
St. Louis, Missouri

Conall M. O'Cleirigh, PhD
University of Miami
Coral Gables, Florida

Michael W. Otto, PhD
Boston University
Boston, Massachusetts

I. Jacqueline Page, PsyD
University of Tennessee Health Science Center
Memphis, Tennessee

Joanne Peart, BA
Emory University
Atlanta, Georgia

David S. Riggs, PhD
University of Pennsylvania School of Medicine
Philadelphia, Pennsylvania

Karen Rowa, PhD
McMaster University
Hamilton, Ontario

Karen B. Schmaling, PhD, ABPP
University of North Carolina at Charlotte
Charlotte, North Carolina

Lori N. Scott, BA
Pennsylvania State University
University Park, Pennsylvania

Alisa R. Singer, MSc
University of Calgary
Calgary, Alberta

Jasper A. J. Smits, PhD
Southern Methodist University
Dallas, Texas

Marika Solhan, BS
University of Missouri-Columbia
Columbia, Missouri

Stephanie D. Stepp, MA
University of Missouri-Columbia
Columbia, Missouri

Eric Stice, PhD
University of Texas at Austin
Austin, Texas

Marilyn J. Strada, MA
University of Nevada, Las Vegas
Las Vegas, Nevada

Michael E. Thase, MD
University of Pittsburgh Medical Center
Pittsburgh, Pennsylvania

Heather Thompson-Brenner, PhD
Boston University
Boston, Massachusetts

Beverly E. Thorn, PhD
University of Alabama
Tuscaloosa, Alabama

Timothy J. Trull, PhD
University of Missouri-Columbia
Columbia, Missouri

Drew Westen, PhD
Emory University
Atlanta, Georgia

PART ONE

GENERAL ISSUES

CHAPTER 1

Diagnosis and Classification

JAMES LANGENBUCHER AND PETER E. NATHAN

A BRIEF OVERVIEW OF CONCEPTUAL UNDERPINNINGS

The craft of psychiatric diagnosis is essential to nearly all clinical, research, and policy endeavors involving mental health. For clinicians, diagnostic systems identify at-risk individuals for prevention services; select other cases for referral and brief treatment; in more serious cases they may suggest special courses of treatment that have been empirically tested; and of course they confer on third-party payers the responsibility to honor charges for that treatment. For scientists, well developed diagnostic systems protect the integrity of human research samples; provide an important heuristic function by suggesting systematic relationships among psychiatric illnesses; allow scientists (and practitioners) from disparate backgrounds to communicate via a consensual nomenclature; and enable epidemiologists to find illness base rates, risk/resilience indicators, and other facts in the data. For policymakers, these rewards of well developed diagnostic systems provide the tools to apportion health and other social resources wisely. But probably most importantly, well developed diagnostic systems provide nothing less than the essential structure for the storage and retrieval of new knowledge as it is gathered in the field (Blashfield & Draguns, 1976), in all ways essential to the scientific enterprise.

Though some diagnostic systems are dimensional or otherwise noncategorical, and will be discussed briefly later, most are categorical or, like *DSM-III, III-R,* and *IV,* "class-quantitative" (Strauss, 1975). Such systems permit additional nuance, such as severity ratings, codes for the presence/absence of special features, and so on, but they require, above all, diagnostic classification. This is so for, as Raven, Berlin, and Breedlove (1971) observed in a seminal monograph in the journal *Science,* "Man is by nature a classifying animal. . . . Indeed, the very development of the human mind seems to have been closely related to the perception of discontinuities in nature" (p. 1210).

Raven and his colleagues used the term *folk taxonomy* to indicate the predisposition of subgroups, especially guildlike groups of craftsmen, to establish categorical nomenclatures (folk taxonomies) for classifying objects in nature that are of special interest to them. Thus, potters have extensive taxonomies of clay, stonecutters of hardness and grain, and so forth. In a classic monograph, the cognitive psychologist Eleanor Rosch (1973) extended this argument by observing that, across human cultures, there are nonarbitrary or "natural" categories that form around perceptually salient *natural prototypes.* Such natural categories could, of course, serve as the basis for the folk taxonomies described by Raven and his coauthors. Rosch explained the key attributes of natural categories: (1) they are nonarbitrary; (2) they are partitioned from continua; (3) they cannot, by use of normal language, be further reduced to simpler attributes; (4) they are easily learned by novices; (5) they serve as natural structures for the organization of more knowledge; and (6) they have indistinct boundaries, encompassing both clear-cut and marginal examples. So, not only do human beings naturally tend to categorize and classify things, as Raven and colleagues argue, Rosch would have it that human beings tend to categorize and classify things in roughly the same way, across cultures and, presumably, across historical eras. It seems a characteristically human thing to do.

In a more recent monograph, Lilienfeld and Marino (1995) extended a Roschian analysis to psychiatric diagnosis, arguing that major psychopathologic entities such as schizophrenia or bipolar illness are, like Roschian or natural prototypes, partitioned from the continuum of human behavior, irreducible to simpler concepts, understood analogously across cultures, have good and bad examples, and so on. This view complements the conceptualization of psychiatric diagnosis as a problem in *prototype categorization* (Cantor, Smith, French, & Mezzick, 1980). Cantor and her colleagues proposed that psychiatric diagnosis follows not a *classic categorization* model (universally accepted criteria, high agreement about class membership, and within-class homogeneity of members) but rather a *prototype categorization* model. Prototype categorization assumes (1) correlated—not necessar-

ily pathognomonic—criteria for class membership, (2) high agreement among classifiers only when classifying cases that demonstrate most of the correlated criteria for class membership (disagreement is expected when cases have a marginal number of category features, or when they bear features from more than one category), and (3) heterogeneity of class membership, because criteria are only correlated, not pathognomonic.

Thus, whereas systems of psychiatric diagnosis have their critics—and many of their arguments will be reviewed later—there is nothing arcane, much less unprecedented, in the actions of a mental health professional who, encountering a new case, lifts a copy of the *DSM* from her desk, matches the properties of the new case to one or more of the *DSM* categories, and then uses the diagnostic result to select treatment, to make a referral, or to rule the case in or out of a research protocol. To the contrary, what the mental health professional is doing is as old, as honored, as universal, and as essentially human as the crafts themselves (Nathan & Langenbucher, 1999).

A BRIEF HISTORY OF DIAGNOSIS

Throughout the classical era, diagnoses were made on the basis of presumed etiology, as when Hippocrates rooted the illnesses he diagnosed (mania, melancholia, and paranoia) in various imbalances of black bile, yellow bile, blood, and phlegm (Zilboorg, 1941). Galen (A.D. 130–210), an influential Greek anatomist who lived more than 500 years later, took much the same view in his descriptions of both normal and abnormal sensations and perceptions as products of a spirit or vapor he called *pneuma psychikon.* Basing diagnostic assessments on such etiologic conceits changed only when the Swiss physician and natural philosopher Paracelsus (1490–1541) developed the concept of *syndromal diagnosis.* Paracelsus defined the syndrome as a group of signs and symptoms that co-occur in a common pattern and thereby, presumably, characterize a particular abnormality or disease state, but for which etiology is unknown, perhaps even unknowable. Syndromal diagnosis is epitomized today in the *DSM,* which continues its focus on the signs and symptoms of diseases, rather than their presumed etiologies, which are unnecessary for diagnostic purposes.

Typically, psychiatric illnesses are organized hierarchically, by the principles of descriptive similarity or shared symptom pictures. Thus, following Paracelsus, more comprehensive and better organized hierarchical classification systems were soon developed, first by Thomas Sydenham (1624–1689), an English physician for whom a childhood chorea is named, and a bit later by the French physician François de Sauvages (1706–1767). Shortly afterward, famed French hospital reformer Phillippe Pinel (1745–1826), pictured in almost every abnormal psychology textbook breaking the chains of the insane in Paris's Bicêtre and Salpêtrière hospitals, proposed a system that included melancholia, mania, mania with delirium, dementia, and idiotism. The appearance of this nomenclature coincided with the development of asylums for the insane, for which Pinel was partly responsible, and certainly contributed to both their humanity and their success. Building on this advance, both Pinel's system and the new availability of large numbers of diagnostically differentiated patients in asylums paved the way for the marked increase in efforts to categorize psychopathology during the nineteenth century.

The victims of serious, chronic psychopathology—what are today understood as organic mental disorders, severe developmental disabilities, dementia, schizophrenia, and bipolar disorder (Nathan, 1998; Spitzer, Williams, & Skodol, 1980)—were permanent residents of these asylums for the mentally ill. The study of their essential features accelerated when the German psychiatrist Karl Kahlbaum (1828–1899) discovered that understanding the premorbid course of *dementia praecox* (which today we call schizophrenia), and the factors that conferred risk for it, helped predict its outcome. The roots of modern syndromal classification, including the *Diagnostic and Statistical Manual of Mental Disorders,* can be traced to Kahlbaum and to fellow German taxonomists Griesinger and Hecker. But no figure in descriptive psychopathology stands taller than Emil Kraepelin (1856–1926), whose successive textbook editions at the end of the nineteenth and beginning of the twentieth centuries anticipated much of what modern-day diagnosticians would find familiar, including detailed medical and psychiatric histories of patients, mental status examination, emphasis on careful observation of signs and symptoms to establish diagnoses, and understanding the psychoses as largely diseases of the brain. Kraepelin's taxonomy of mental illness has a strikingly contemporary feel and includes many of the terms used today.

In the twentieth century, more and more mental health practice took place outside the mental asylums, to encompass the military services, private clinics and office practice, company-supported mental health and substance abuse services, and educational institutions at all levels. As a result, nosologies grew broader and increasingly complex in instruments published by the National Commission on Mental Hygiene/Committee on Statistics of the American Medico–Psychological Association in 1917 and the American Psychiatric Association/New York Academy of Medicine (1933). This was both fortunate and necessary, for during World

War II, unexpectedly, most psychological casualties resulted from nonpsychotic, acute disorders like substance abuse, depression, and the anxiety disorders, with extraordinarily high base rates among combat personnel. Clearly, the impact of these conditions on the war effort required development of a nomenclature that provided substantially greater coverage of these conditions so that they could be accurately identified, treated, and their sufferers returned to service.

A COMMON U.S. NOMENCLATURE: *DSM-I* AND *DSM-II*

Although the U.S. War Department worked hard to develop such a system in response to the flood of wartime psychiatric casualties, it was only in 1946 that representatives of the Veterans Administration, the War Department, and the civilian mental health community led by the American Psychiatric Association (APA) began to consider how to create a nomenclature that would meet their diverse needs. Their efforts led to the publication, in 1952, of the first edition of the APA's *Diagnostic and Statistical Manual of Mental Disorders (DSM-I).*

The *DSM-I* (APA, 1952) was the first comprehensive syndromal system developed. As such, it was designed to offer mental health professionals a common diagnostic language through which to communicate about their patients and their research findings. Its appearance sparked a similar effort in Europe that ultimately caused the World Health Organization (WHO) to add a mental disorders section to the eighth edition of the *International Classification of Diseases* (*ICD-8;* WHO, 1967). Despite its promise, *DSM-I* (and *DSM-II* [APA, 1968], which closely resembled it) shared serious problems that markedly compromised their diagnostic reliability, validity, and utility.

Most obviously, the manuals contained relatively little textual material: The *DSM-I* contained 130 pages and fewer than 35,000 words; *DSM-II* was a mere four pages longer. As a consequence, these early efforts provided only brief descriptions of each syndrome, insufficient for reliable diagnoses. Moreover, the signs and symptoms of each syndrome were not empirically based. Instead, they represented the accumulated clinical wisdom of the small number of senior academic psychiatrists who staffed the *DSM* task forces. As a result, the diagnostic signs and symptoms that interested task force members were imperfectly related to the clinical experiences of mental health professionals working in public mental hospitals, mental health centers, and the like. Consequently, clinicians very often failed to agree with one another when assigning diagnoses based on *DSM-I* and *DSM-II,* whether they were presented with the same diagnostic information (*interclinician agreement;* Beck, Ward, Mendelson, Mock, & Erbaugh, 1962; Nathan, Andberg, Behan, & Patch, 1969) or they reevaluated the same patient after a period of time had passed (*diagnostic consistency;* Zubin, 1967).

Not surprisingly, the low reliability of *DSM-I* and *DSM-II* diagnoses affected both their validity and clinical utility. If clinicians could not agree on a diagnosis, they were unlikely to be able to validate it against other measures (Black, 1971), to have confidence in predictions of the future course of diagnosed disorders (Nathan, 1967), or to create the diagnostically homogeneous groups of patients necessary to spur substantive advances in etiological or treatment research (Nathan & Harris, 1980).

Just as predictably, the low reliability and validity of *DSM-I* and *DSM-II* diagnoses raised ethical concerns among practitioners and scholars. Psychiatrist Thomas Szasz (1960) created a national furor over what he considered the dehumanizing, stigmatizing consequences of psychiatric "labeling," ultimately concluding that the modern categories of psychiatric illness were mere "myths." Szasz's ideas gained empirical substance in 1973 when psychologist David Rosenhan published, in the world's most prestigious journal, *Science,* one of the most widely cited studies in psychiatry, "On Being Sane in Insane Places." At Rosenhan's behest, eight peers, friends, and graduate students presented for treatment to various psychiatric hospitals in northern California, complaining of "hearing voices." Auditory hallucinations are, of course, a "first-rank" symptom of schizophrenia (Schneider, 1959), and all eight pseudopatients were admitted to hospital. Immediately thereafter, they stopped complaining of the voices and denied any other symptoms of psychosis. Nonetheless, all were diagnosed as psychotic, and their subsequent behavior was construed in light of that label. Quite normal reactions they manifested, such as being wary of strange and perhaps menacing fellow patients, were characterized in chart notes and staff meetings as the products of paranoid and delusional processes. Summarizing his findings, Rosenhan concluded, "The normal are not detectably sane" (1973, p. 252), a damning assertion indeed. Clearly, psychiatric diagnosis had come as far as it possibly could as an "art" practiced in an arcane fashion by an elite group of the initiated. The time was ripe for its transformation into a science.

EMERGENCE OF THE NEO-KRAEPELINIAN TRADITION: *DSM-III* AND *DSM-III-R*

Antecedents

Beginning in the late 1960s, psychiatrist Robert Spitzer and colleagues at the New York State Psychiatric Institute devel-

oped several structured diagnostic interviews, including the Mental Status Schedule (Spitzer, Fleiss, Endicott, & Cohen, 1967) and the Psychiatric Status Schedule (Spitzer, Endicott, Fleiss, & Cohen, 1970), in an effort to begin to gather empirical data on diagnostic syndromes. Spitzer and his colleagues also developed computer programs called DIAGNO and DIAGNO-II that were designed to use the syndromal information gathered by the Mental Status Schedule to assign more reliable clinical diagnoses (Spitzer & Endicott, 1968, 1969).

Sharing a similar commitment to developing an empirically based, more reliable diagnostic system, researchers at Washington University in Saint Louis published an important article in 1972 (Feighner et al., 1972) that set forth explicit diagnostic criteria—the so-called Feighner criteria—for 16 major disorders. Their intent was to replace the vague and unreliable descriptions of *DSM-I* and *DSM-II* with systematically organized, empirically based diagnostic criteria, helping researchers to establish the diagnostically homogeneous and predictively valid experimental groups for which they had long striven in vain. The format of the Feighner criteria greatly influenced the format for diagnostic criteria adopted in *DSM-III.* A derivative of Feighner's work, the Research Diagnostic Criteria (RDC), developed jointly by the New York State Psychiatric Institute and Washington University groups (Spitzer, Endicott, & Robins, 1975), was published in 1975. Designed to permit empirical testing of the presumably greater reliability and validity of the Feighner criteria, the RDC criteria yielded substantially greater diagnostic reliability than the equivalent *DSM-II* disorders (Helzer, Clayton, et al., 1977; Helzer, Robins, et al., 1977), and so constituted a great step forward.

This work, rooted in the idea of psychiatric diagnosis as a rigorously developed and universally applied scientific tool, defined what came to be known as the *neo-Kraepelinian* school of U.S. psychiatry (Blashfield, 1984). Drawing largely from the groups that formulated the RDC—psychiatry faculty at the Washington University School of Medicine in Saint Louis and the Columbia University College of Physicians and Surgeons in New York—neo-Kraepelinian diagnostic research during the 1970s laid the groundwork for the revolutionary advances of *DSM-III.* Like Kraepelin himself, the neo-Kraepelinians endorsed the existence of a boundary between "pathological functioning" and "problems in living," viewed mental illness as the purview of medicine, and believed in the importance of applying the scientific method so that the etiology, course, prognosis, morbidity, associated features, family dynamics, predisposing features, and treatment of psychiatric illnesses could be elucidated more clearly.

Diagnostic Criteria

Five years after the RDC criteria were published, *DSM-III* appeared (APA, 1980), heralding substantial advances in the reliability, validity, and utility of syndromal diagnosis. Based in large part on the RDC, the inclusion in *DSM-III* of rigorously designed *diagnostic criteria* and, in an appendix, *diagnostic decision trees,* represented the new instrument's most significant advance. The criteria were designed to organize each syndrome's distinguishing signs and symptoms within a consistent format—they were, in scientific parlance, *operationalized,* so that each clinician who used them would define each sign and symptom the same way, and process the resulting diagnostic information in a consistent manner. This degree of detail in the diagnostic information available to *DSM-III*'s users contrasted sharply with the paucity of such detail in *DSM-I* and *DSM-II.*

Several structured and semistructured diagnostic interviews based on the *DSM-III,* very distant descendants of the Mental Status Schedule and the Psychiatric Status Schedule, were published around the time *DSM-III* appeared, in a related effort to enhance diagnostic reliability and, especially, to spur research. The best known of these was the NIMH Diagnostic Interview Schedule (DIS; Robins, Helzer, Croughan, & Ratcliff, 1981), a structured interview designed for nonclinician interviewers. The semistructured Structured Clinical Interview for *DSM-III* (SCID; Spitzer, 1983; Spitzer & Williams, 1986), designed for use by clinicians, was also published around the same time. These important, and in most ways unprecedented, new instruments provided the data-gathering structure both for major new epidemiologic efforts (e.g., Epidemiologic Catchment Area study [Regier et al., 1984], National Comorbidity Survey [e.g., Kessler, Sonnega, Bromet, Hughes, & Nelson, 1995; Kessler, Stein, & Berglund, 1998]) and for a host of clinical and preclinical studies, because they insured the internal validity of the research by helping ensure that the samples of human psychopathology were well characterized diagnostically. *DSM-III-R,* published in 1987, was a selective revision of *DSM-III* that retained the advances of the 1980 instrument and incorporated generally modest changes in diagnostic criteria that new clinical research (to a great extent dependent on findings produced by the application of the DIS and SCID to human research samples) suggested should be a part of the diagnostic system. It was in this way that diagnostic research "bootstrapped" its way from the dismal days of Rosenhan to the well-regarded science it is today, and its products, although not universally successful, have been impressive indeed.

Utility and Validity

DSM-III and *DSM-III-R* addressed their predecessors' disappointing diagnostic validity and utility in several ways (Spitzer et al., 1980). To begin with, both volumes are much larger than their predecessors, in part to accommodate inclusion of more than three times as many diagnoses, in part to provide detailed information on each syndrome along with its defining diagnostic criteria. The expansion of syndrome descriptions made it easier for clinicians to describe more precisely their patients' behavior, and to understand their suffering in the context of their milieu.

Another advantage of *DSM-III* and *DSM-III-R* was that they assessed patients along five dimensions, or axes: Psychopathology was diagnosed on Axes I and II; medical conditions impacting on the mental disorders were noted on Axis III; the severity of psychosocial stressors affecting the patient's behavior was noted on Axis IV; and the patient's highest level of adaptive functioning was noted on Axis V. The additional information available from multiaxial diagnosis was presumed to be more useful for treatment planning and disposition than the single diagnostic label available from *DSM-I* and *DSM-II.*

Reliability and Stability

A very substantial number of reliability studies of the *DSM-III* and *DSM-III-R* diagnostic criteria were published. Almost without exception, they pointed to much greater diagnostic stability and interrater agreement for these instruments than for their predecessors, *DSM-I* and *DSM-II.* Enhanced reliability was especially notable for the diagnostic categories of schizophrenia, bipolar disorder, major depressive disorder, and substance abuse and dependence; the reliability of the personality disorders, some of the disorders of childhood and adolescence, and some of the anxiety disorders has been less encouraging (e.g., Fennig et al., 1994; Klein, Ouimette, Kelly, Ferro, & Riso, 1994; Mattanah, Becker, Levy, Edell, & McGlashan, 1995), but this has been due to a variety of reasons, including conceptual underspecification (in the case of the personality disorders), and the inherently transitory of self-correcting nature (diagnostic stability problems) of some others (disorders of childhood and adolescence and some forms of anxiety).

Thus, despite these explicit efforts to enhance the diagnostic utility and validity of *DSM-III* and *DSM-III-R,* it did not prove easy to document the impact of these efforts. The absence of documented etiological mechanisms, with associated laboratory findings, by which the diagnoses of many physical disorders are confirmed—the "gold standard"—made establishing the construct validity of many *DSM-III* and *DSM-III-R* diagnoses difficult (Faraone & Tsuang, 1994). As noted later in this chapter, the same problem continues to stand in the way of attempts to validate *DSM-IV* diagnoses.

Although the *DSM-III* and *DSM-III-R* diagnostic criteria enhanced the instruments' diagnostic reliability, diagnostic stability continued to be an issue for diagnosticians because of changes in patient functioning over time. Thus, in a study of the six-month stability of *DSM-III-R* diagnoses in first-admission patients with psychosis, Fennig et al. (1994) reported that whereas affective psychosis and schizophrenic disorders showed substantial diagnostic stability, stability for subtypes of these conditions was less stable. Changes in patient functioning were seen as responsible for 43 percent of these diagnostic changes. In like fashion, Nelson and Rice (1997) reported that the one-year stability of *DSM-III* lifetime diagnoses of obsessive-compulsive disorder (OCD) turned out to be surprisingly poor: Of OCD subjects in the ECA sample they followed, only 19 percent reported symptoms a year later that met the OCD criteria. Mattanah and his colleagues (1995) reported that the diagnostic stability of several *DSM-III-R* disorders was lower for a group of adolescents two years after hospitalization than for the same diagnoses given adults. These and similar studies of diagnostic stability emphasized the extent to which diagnostic reliability is dependent not only on the validity of diagnostic criteria but on the inherent symptom variability of disorders over time as well.

Also, researchers using *DSM-III* and *DSM-III-R* diagnostic criteria undertook research during the years following their appearance to validate several of the manual's major diagnostic categories, including schizophrenia and major depressive disorder, despite the absence of a gold-standard criterion of validity. Our brief mention of validation studies includes only Kendler's familial aggregation and coaggregation research findings, both because they represent a particularly powerful approach to validation and because the findings generally mirror those found by others, but many others could be adduced.

When Kendler, Neale, and Walsh (1995) examined the familial aggregation and coaggregation of five hierarchically defined disorders—schizophrenia, schizoaffective disorder, schizotypal/paranoid personality disorder, other nonaffective psychoses, and psychotic affective illness—in siblings, parents, and relatives of index and comparison probands, they reported that although schizophrenia and psychotic affective illness could be clearly assigned to the two extremes of the schizophrenia spectrum, the proper placement of schizoaf-

fective disorder, schizotypal/paranoid personality disorder, and other nonaffective psychoses could not be clearly made. In a companion report, Kendler and his coworkers (1995) found that probands with schizoaffective disorder differed significantly from those with schizophrenia or affective illness in lifetime psychotic symptoms as well as outcome and negative symptoms assessed at follow-up. Moreover, relatives of probands with schizoaffective disorder had significantly higher rates of schizophrenia than relatives of probands with affective illness.

Although Kendler's family research method validated only a portion of schizophrenic spectrum disorder diagnoses, he and his colleagues (Kendler et al., 1996; Kendler & Roy, 1995) were able by the same methods to strongly support the validity of the *DSM-III* major depression diagnostic syndrome. However, when Haslam and Beck (1994) tested the content and latent structure of five proposed subtypes of major depression, clear evidence for discreteness was found only for the endogenous subtype; the other proposed forms lacked internal cohesion or were more consistent with a continuous or dimensional account of major depression.

Criticisms

Although *DSM-III* and *DSM-III-R* represented major advances, they were widely criticized. This was particularly so for *DSM-III,* the first manual to truly shatter the mold in which prior nosologies had been cast. One major source of concern was that *DSM-III* incorporated more than three times the number of diagnostic labels in *DSM I.* Prominent clinical child psychologist Norman Garmezy (1978) expressed the concern that this proliferation of diagnostic labels would tempt clinicians to pathologize unusual but normal behaviors of childhood and adolescence, a criticism more recently directed at *DSM-IV* (Houts, 2002). In a similar vein, social workers Kirk and Kutchins (1992) accused the instrument's developers of inappropriately labeling "insomnia, worrying, restlessness, getting drunk, seeking approval, reacting to criticism, feeling sad, and bearing grudges . . . [as] possible signs of a psychiatric illness" (p. 12).

Thus, the definition of mental disorder developed for *DSM-III* (and retained in *DSM-III-R* and *DSM-IV*) has been criticized for being both too broad and encompassing of behaviors not necessarily pathological, and for offering poor guidance to clinicians who must distinguish between uncommon or unusual behavior and psychopathological behavior. Addressing these concerns, Spitzer and Williams (1982) defended the *DSM-III* approach (and by extension, the entire ensuing *DSM* tradition) by noting that the intention of the framers was to construct a nomenclature that would cast as wide a clinical net as possible, in order that persons suffering from even moderately disabling or distressing conditions would receive the help they needed.

But *overdiagnosis* was not the only rifle leveled at the *DSM* tradition. Schacht and Nathan (1977), Schacht (1985), and others questioned the frequent emphasis in *DSM-III* on disordered brain mechanisms in its discussions of etiology, as well as its apparent endorsement of pharmacological treatments in preference to psychosocial treatments for many disorders. In response, Spitzer (1983) noted that the *DSM-III* text simply reflected the state of knowledge of etiology and treatment. Similar concerns have been voiced about *DSM-IV* by Nathan and Langenbucher (1999).

DSM-III and its successors have also been criticized for their intentionally atheoretical, descriptive position on etiology. In a debate on these and related issues (Klerman, Vaillant, Spitzer, & Michels, 1984), these critics charged that an atheoretical stance ignored the contributions of psychodynamic theory to a fuller understanding of the pathogenesis of mental disorders, as well as to the relationship between emotional conflict and the ego's mechanisms of defense. But in the same debate, Spitzer questioned the empirical basis for the claim that psychodynamic theory had established the pathogenesis of many of the mental disorders. Clearly, these are matters on which much has still to be written and argued, as it surely will be.

THE PRESENT EMPIRICALLY BASED NOMENCLATURE: *DSM-IV*

The *DSM-IV* Process

The principal goal of the *DSM-IV* process was to create an empirically based nomenclature that improved in important ways on *DSM-III* and *DSM-III-R* (Frances, Widiger, & Pincus, 1989; Nathan, 1998; Nathan & Langenbucher, 1999; Widiger & Trull, 1993). To achieve this goal, a three-stage process was used. The process began with the appointment of 13 four- to six-person work groups of experts on the major diagnostic categories. Each work group began by undertaking systematic literature reviews designed to address unresolved diagnostic questions. When the literature reviews failed to resolve them, the work groups sought clinical data that might be capable of casting more light on outstanding questions; 36 reanalyses of existing patient data sets were ultimately completed. The work groups also designed and carried out 12 large-scale field trials involving more than seven thousand participants at more than 70 sites worldwide.

The *DSM-IV* development process is thoroughly chronicled in four *Sourcebooks* edited by Thomas Widiger and colleagues (Widiger et al., 1994, 1996, 1997, 1998). There, important literature reviews are archived and findings from data reanalyses and field trials are summarized. Most of the field trials contrasted the diagnostic sensitivity and specificity of alternative sets of existing diagnostic criteria, including those of *ICD-10, DSM-III-R,* and *DSM-III,* with one or more sets of new criteria, the *DSM-IV Options.* Many of the "options" explored the impact on diagnostic reliability of changes in the wording of criteria or the minimum number required to meet *diagnostic threshold* and permit formal diagnoses to be made.

Reliability and Validity

Most of the early data on the reliability and validity of *DSM-IV* diagnoses came from the field trials. Generally, the data suggested modest increases in the reliability of a few diagnostic categories (e.g., oppositional defiant disorder and conduct disorder in children and adolescents, substance abuse and dependence) and validity (e.g., autistic disorder; oppositional defiant disorder in childhood and adolescence). Unfortunately, they also reported no real progress in addressing the substantial reliability problems of the personality disorders, the sleep disorders, the disorders of childhood and adolescence, and some of the disorders within the schizophrenic spectrum. These continue to be thorny problems that scientists developing *DSM-V* are now hard at work to solve.

Because estimates of diagnostic reliability reflect, at least in part, the stability of the disorder's symptoms, a number of studies of *DSM-IV* symptom stability have been undertaken. To this end, Koenigsberg and his colleagues (2002) explored the instability of affective symptoms in borderline personality disorder; Mataix-Cols and his coworkers studied symptom stability in adult obsessive-compulsive disorder (2002), and Shea and her fellow investigators (2002) explored the short-term diagnostic stability of schizotypal, borderline, avoidant, and obsessive-compulsive personality disorders. It is not accidental that in all three instances the temporal stability of disorders with typically unstable symptom patterns was studied. Interestingly, in all three studies, the investigators found greater than anticipated symptom stability.

Of the relatively few very recent reliability studies appearing in the literature—their paucity reflects the substantial number of such studies already published in the *Sourcebooks,* as well as agreement among scholars and researchers that diagnostic reliability for most *DSM-IV* diagnostic categories is satisfactory—one reported good interrater agreement among experienced psychiatrists for *DSM-IV* diagnoses of bipolar II disorder (Simpson et al., 2002), whereas a second found "good to excellent reliability" for the majority of current and lifetime *DSM-IV* anxiety and mood disorder diagnoses (Brown, Di Nardo, Lehman, & Campbell, 2001). In addition, four recent pieces attested to the predictive validity of a diverse group of *DSM-IV* disorders. Kim-Cohen and her colleagues (2003) reported that between 25 percent and 60 percent of a large sample of British adults with a current psychiatric diagnosis had a history of conduct and/or oppositional defiant disorder, making the latter particularly predictive of adult disorder. Assessing psychiatric disorders in a random community sample of U.S. adolescents, Johnson, Cohen, Kotler, Kasen, and Brook (2002) found that depressive disorders during early adolescence were associated with elevated risk for the onset of eating disorders, and disruptive and personality disorders were independently associated with elevated risk for specific eating or weight problems. Yen and her colleagues (2003) reported that the diagnosis of borderline personality disorder (BPD) among patients in a more variegated group, when combined with a history of drug use, significantly predicted suicide attempts during a two-year follow-up; when BPD was controlled, a worsening in the course of major depressive disorder and of substance use disorders in the month preceding the suicide attempt was also a significant predictor of suicide. Following the five-year clinical course of almost 600 men and women with diagnoses of alcohol abuse or alcohol dependence, Schuckit and colleagues (2001) observed that the *DSM-IV* diagnosis of alcohol dependence predicted a chronic disorder with a relatively severe course, whereas *DSM-IV* alcohol abuse predicted a less persistent, milder disorder that did not usually progress to dependence. Results from these four studies support a growing consensus that the enhanced reliability of a number of *DSM-IV* diagnoses, reflecting more accurate diagnostic criteria, has led to their greater predictive validity.

In a thoughtful recent article that asked clinicians to distinguish between the concepts of diagnostic validity and diagnostic utility that also has relevance to the continuing dimensional/categorical controversy, Kendell and Jablensky (2003) observed that, despite historical assumptions to the contrary, little evidence demonstrates that most currently recognized mental disorders are separated by natural boundaries. Although these authors make the case that diagnostic syndromes should be regarded as valid only if they have been shown to be discrete entities with natural boundaries, they make a strong case for believing that many of these entities nonetheless possess high utility because they do provide valuable information on outcome, treatment response, and etiology. That is, with reference to the points with which we began this chapter, Kendell and Jablensky believe that *DSM-*

IV diagnoses are useful folk taxonomies in the sense described by Raven and colleagues, even if they do not meet the standard of natural prototypes described by Eleanor Rosch.

Gender and Cultural Bias

In response to the controversy surrounding *DSM-III-R*'s estimates that more women than men merit the diagnoses of histrionic PD and dependent PD, the *DSM-IV* text now avoids specifying gender prevalence rates for these disorders. *DSM-IV* has also added three PDs (schizoid, schizotypal, and narcissistic) to the three (paranoid, antisocial, and obsessive-compulsive) that *DSM-III-R* indicated were diagnosed more often in males than in females. In reviewing these changes, Corbitt and Widiger (1995) asked whether *DSM-IV* has unintentionally introduced diagnostic bias, in a laudable effort to combat it, by going beyond the modest empirical data on gender prevalence rates for the histrionic and dependent PDs.

Two recent studies examined the impact of ethnicity on rates of psychiatric disorders, in a continuing expression of interest in ethnicity and diagnosis stimulated in part by *DSM-IV*'s Appendix I. Minsky, Vega, Miskimen, Cara, and Escobar (2003) reported significantly higher rates of major depression for Latinos in a survey of differential diagnostic patterns among Latino, African American, and European American psychiatric patients drawn from a large behavioral health service delivery system in New Jersey. However, these authors were unsatisfied with the range of possible explanations for this unexpected finding. Canino and her colleagues (2004) examined rates of child and adolescent disorders in Puerto Rico, finding prevalence rates "that were generally comparable with those found in previous surveys" and broadly in line with previous surveys of children and adolescents on the U.S. mainland.

Criticisms

Although there is widespread agreement about the enhanced empirical base that underlies *DSM-IV,* many persons involved in the development of the instrument acknowledge limitations on full utilization of the extensive empirical database because of unavoidable, biased or misleading interpretations of the data (e.g., Kendler, 1990; Widiger & Trull, 1993). Responding to related criticisms that professional issues overshadowed scientific ones in the creation of *DSM-IV* (e.g., Caplan, 1991; Carson, 1991; Kirk & Kutchins, 1992), Widiger and Trull (1993) defended attention by the drafters of *DSM-IV* to issues of utility that sometimes preempted issues of validity, as when a valid diagnosis was de-emphasized because so few patients met its criteria. Nonetheless, even though the *DSM-IV* Task Force had to be sensitive to a variety of forensic, social, international, and public health issues, Widiger and Trull described the result as largely an empirically driven instrument. The *DSM* tradition, and the much enhanced approach to diagnostic inquiry it helped promulgate, has had impressive impact on how scientists conduct research and, thus, on how clinicians approach their patients.

TWO CRITICAL CASES OF DIAGNOSIS

Epidemiology: The CIDI in the NCS

Prior to the arrival in 1980 of a rule-guided diagnostic system, *DSM-III,* the basic fact of mental illness was appreciated, and some preliminary studies in psychiatric epidemiology—such as the New Haven Study (Hollingshead & Redlich, 1955) and the Midtown Manhattan Study (Srole, Langer, Michael, Opler, & Rennie, 1962)—were conducted. However, not even the best-informed scientist of the time knew much about how prevalent mental illnesses were, how they co-occurred, how they were concentrated in certain age ranges, what factors seemed to predispose to their presence and absence (e.g., risk and resiliency factors), and so on. Firm findings require reliable diagnoses, and these were impossible in the absence of rule-guided diagnostic systems like *DSM-III.* This gap in diagnostic methodology made investment in large-scale epidemiologic research by the U.S. government unattractive. Consequently, because health policymakers had little basis on which to make informed judgments, groups of underidentified persons affected by psychiatric conditions—PTSD patients, patients with mild depression, children with learning disorders, and so on—may well have suffered needlessly.

This situation changed with the development of the Epidemiologic Catchment Area study, which deployed an important tool, the structured diagnostic interview, for the first time in a large-scale epidemiologic study. The ECA involved face-to-face interviews of a stratified sample of more than 18,000 adult community respondents in five states during the early 1980s. Its goal was to establish the prevalence of a very wide range of mental and substance use disorders in the United States. This goal became possible with the development of the Diagnostic Interview Schedule, based on *DSM-III* diagnostic categories, and structured to permit specially trained non-mental-health professionals to interview and diagnose respondents reliably. The DIS was an early example of what became, during the late 1980s and 1990s, a large and sophisticated family of fully and semistructured or "guided" diagnostic interviews—including the CIDI (Robins et al.,

1988), SCID (Spitzer, Williams, Gibbon, & First, 1992), SADS (Endicott & Spitzer, 1978), PRISM (Hasin et al., 1996), and many others—developed by numerous independent research teams to facilitate both clinical and epidemiological research. It is true that most of these instruments require extensive interviewer training and can take several hours to complete, but they are designed to do something that was never before possible: to yield full knowledge of that respondent's psychiatric state and past history, formal diagnoses of illnesses that meet diagnostic threshold, and even the presence of individual symptom, symptom severity, symptom onset and offset patterns, and subclinical states.

Between September 1990 and February 1992, using a descendant of the DIS called the CIDI-UM, Kessler and colleagues undertook the successor to the ECA, the National Comorbidity Study (NCS). The NCS gathered data on demographics, psychiatric and health functioning, quality of life, and many other domains from a stratified national sample of more than 8,000 Americans aged 15 to 54 years. Like the ECA before it, the NCS dataset has generated scores of important epidemiologic and descriptive studies on issues as diverse as adolescent depression (Kessler, Avenevoli, & Merikangas, 2001), generalized anxiety disorder (Wittchen, Zhao, Kessler, & Eaton, 1994), symptom progression of alcohol dependence (Nelson, Little, Heath, & Kessler, 1996), and many others. Because of such studies, we now understand, within the limitations of our current concepts, how prevalent psychiatric illnesses are, how they onset, what genetic and other factors seem to predispose to them, and many other matters of crucial public concern. In fact, data derived from the NCS—and thus directly derived from the *DSM-IV* and the structured interviews it made possible—are constantly appealed to by mental health administrators and policymakers charged with assessing and predicting service and research requirements prior to distributing resources that are increasingly scarce and hard to come by.

Treatment Research: SCID in Randomized Trials

Prior to the 1980s, researchers charged with the design of clinical trials—say, studies of patients with recurrent major depression who could be used in the test of a new medication—suffered as a group from two important limitations: First, without a fairly long conversation with the doctor or diagnostic technician who admitted the subject to trial, no independent observer could have confidence that any particular case indeed met criteria for major depression as developed in *DSM-II;* and, second, no independent observer could have confidence that any particular case did not meet criteria for other psychiatric illnesses in addition to the illness of interest, perhaps much more serious ones, ones that would "wash out" the effects of treatment on the less severe illness. Subjects in such trials were typically deemed eligible for research on the basis of clinician judgment, chart review, and perhaps some narrow-band assessments directed at scaling the severity of the illness of interest, such as the Beck Depression Inventory (BDI; Beck, Ward, Mendelson, Mock, & Erbaugh, 1961), the State-Trait Anxiety Inventory (Speilberger, Gorsuch, & Lushene, 1970), and many others. It is inevitable that some, perhaps many, such cases did not suffer from the illness of interest at all and that a great many other such cases suffered from additional illnesses that confounded the results of the study. To use an analogy, it was as though chemists were charged with developing a new line of anticancer drugs while being blind to the identity of the powders and fluids on their workbenches and unsure of the illnesses from which their experimental subjects suffered.

But just as epidemiologic research was transformed by the availability of a host of fully structured and semistructured or "guided" diagnostic interviews in the 1980s, so, at the same time, did these same diagnostic interviews transform clinical research. Beginning in the 1980s with the development of the SCID, in particular, and continuing more vigorously now, editorial opinion governing the publication of clinical research involving psychiatric groups has required the administration of guided diagnostic interviews to prospective participants in order to protect the integrity of samples. It is difficult—impossible, in most venues—to publish treatment research results when participants have not been "SCID-ed," that is, thoroughly interviewed prior to trial enrollment by skilled diagnosticians using a diagnostic interview like the SCID or one of its close cousins. Is this without cost? No. Initial, preenrollment assessments regularly require hours, even days, to thoroughly characterize the prospective enrollee's history and current clinical state. But, is it worth it? We need cite nothing more than the recent development of *parity* between medical and mental health coverage (Goldman, Rye, & Sirovatka, 1999), a result based on the greater respect accorded research findings in the latter field to assert that it is surely so.

FUTURE DIRECTIONS

Although theoretical and methodological advances have driven forward much of the very clear advantages of *DSM-IV* over its predecessors, application of still other, emergent, research techniques are poised to do much to aid the understanding of, not only what psychiatric illnesses look like, but how they develop over time, what are their essential versus

nonessential characteristics, how they might be further split into meaningful subgroups, and so on. Following, a few of the more promising techniques are briefly reviewed.

New Research Tools

As reviewed previously, the 1970s saw the emergence of an empirical, atheoretical approach to psychiatric diagnosis that blossomed in the work of the neo-Kraepelinian school of U.S. psychiatry. The neo-Kraepelinian movement argued that psychiatric diagnosis, like any branch of medicine, should be based solidly on empirical research (Compton & Guze, 1995). To fulfill this demand, and seeking a strong methodological and empirical base, diagnostic research in the 1970s came to borrow heavily from classical test theory (CTT), whereby such parameters as reliability, internal consistency, and predictive power grew in interest (Baldessarini, Finkelstein, & Arana, 1983). Precisely because of the nature of the methodological problems inherent in diagnostic research, researchers have been required to develop or import from other fields empirical approaches as well, such as epidemiologic concepts and methods, advanced quantitative approaches, and others.

The neo-Kraepelinians formed the core group whose work resulted in the *DSM-III* and later versions of the *DSM*. Their early and important contribution to diagnostic validation models (e.g., Robins & Guze, 1970) is one of the most widely cited papers in psychiatry. The Robins and Guze validation model proposed testing or *validating* diagnostic categories against five criteria: (1) clinical description (the degree to which the symptoms of the disorder cohere and logically connect); (2) laboratory studies (the degree to which the disorder can be seen to covary with physiological markers); (3) delimitation from other disorders (the degree to which the disorder can be distinguished from others, even though some features may overlap); (4) follow-up studies (the degree to which the disorder is stable across time); and (5) family studies (the degree of heritability of the disorder). To this basic model, Andreasen (1995), believing that psychiatry's neuroscience base is key to its future, added neurophysiological and neurogenetic tests. Contemporary validation efforts deploy a mix of clinical, epidemiological, genetic-familial, and neurobiological strategies, some of which will be reviewed in the following sections.

To fulfill the research needs of validation models such as this, a number of powerful advances in quantitative methods sharpened nosological research in the past quarter century. These included both traditional exploratory as well as confirmatory factor analysis (CFA; Cole, 1987; Jöreskog & Sörbom, 1989), to study the presumed internal coherence or unidimensionality of criterion arrays for such diagnoses as borderline personality (e.g., Sanislow et al., 2002) or somatization (Robbins, Kirmayer, & Hemami, 1997); cluster analysis (e.g., Ward, 1963), to discover in the data naturally occurring groups of respondents who may represent *subtype* manifestations of such disorders as mania (Dilsaver, Chen, Shoaib, & Swann, 1997), schizophrenia (Dollfus et al., 1996), or personality disorders (Morey, 1988); receiver/operator characteristic analysis (ROC; Murphy, Berwick, Weinstein, & Borus, 1987), to correct a shortcoming of the *DSM* tradition (the promulgation of clinical thresholds or "cut-points" for formal diagnosis arrived at by expert consensus rather than by quantitative means) by plotting sensitivity against specificity, thus suggesting optimally balanced diagnostic thresholds for such disorders as mania (Cassidy & Carroll, 2001), ADHD (Mota & Schachar, 2000) or traumatic grief (Prigerson et al., 1999), and others. Additional advanced quantitative methods will be reviewed in the following sections. The efforts of all have borne concrete benefits in many areas of research and service delivery.

Latent Class Analysis

Latent class analysis (LCA; McCutcheon, 1987) is a multivariate method that, like the earlier method of cluster analysis, finds structural relationships between cases in a dataset as a function of their status on a set of *manifest variables.* The assumption of users of LCA is that the manifest variables ". . . are imperfect indicators of an underlying latent variable with a finite number of mutually exclusive classes" (Peralta & Cuesta, 2002, p. 415). Whereas *latent profile analysis* is a variation on LCA in which continuous rather than categorical variables are used, LCA itself uses categorical variables, either Likert-type scores or, more commonly, binomial variables. As such, LCA is ideally suited to diagnostic research, where binomial variables—symptom present/absent—are of critical import, and is reviewed here.

Latent classes (they are referred to as *latent* because they are not directly observed, but inferred from the status of groups of cases on the *manifest variables*) fully structure the cases in a dataset with respect to the manifest variables. LCA uses maximum-likelihood estimates in an iterative way to produce model parameters, such as the number of latent classes, or the proportion of cases that fall into each latent class, that best model, or account for, the observed relationships between cases and between manifest variables. An advantage over older cluster analytic techniques is that LCA finds the ideal number of latent classes by testing goodness-of-fit for models with increasing numbers of classes, with the minimum number of latent classes still showing a significant fit by likelihood ratio chi-square (and other methods) being,

in most cases, the preferred solution. LCA also produces, for each case, the probability of its proper fit within each latent class, though cases are classified into the class for which they have the greatest a posteriori probability of membership.

LCA has been of increasing interest to scientists interested in diagnostic issues, such as subtyping research. For instance, Peralta and Cuesta (2003) applied LCA to the abiding problem of psychotic disorders, which, as has been shown previously, have sometimes been difficult to parse adequately. With access to data on 660 psychiatric inpatients diagnosed by both *DSM-IV* and *ICD-10* criteria, Peralta and Cuesta entered 16 variables into the LCA procedure, including data on 15 symptoms (e.g., delusions, manic symptoms, acute versus gradual onset) and whether the data were derived from lifetime versus current (index episode) examination. The study's aims were actually quite ambitious, including a look at the concordance of *DSM* versus *ICD* diagnoses, to examine the relationship between case classifications made by *DSM/ICD* versus empirical LCA, and others. But a main intent was to determine, on the grounds of naïve empiricism, exactly how a large and heterogeneous group of mixed psychotic patients fall naturally into categories, irrespective of how *DSM* or *ICD* would classify them.

Though the patients as a group carried any of eight *DSM* diagnoses (any of seven, when diagnosed by *ICD*), the authors found that a fewer number—five categories—best modeled the data. These were (1) schizophrenia, which was characterized by disorganization, lack of insight, "negative" or "deficit" symptoms (e.g., poverty of speech, avolition), and residual symptoms; (2) psychosis, marked by delusions, hallucinations and lack of insight; (3) schizomanic/schizobipolar disorder, marked by disorganization, lack of insight, moderate depression, and acute onset; (4) schizodepression, with prominent negative, depressive, and residual symptoms; and (5) mixed psychosis, featuring typically moderate levels of many mixed symptoms. These results stand in marked contrast to the structure of *DSM* and *ICD,* with, historically, a fairly pronounced distinction, at the psychotic level of functioning, between illnesses featuring disorganized thought versus unregulated affect, and this area requires much more study. But it is fair to say that studies such as Peralta and Cuesta, using LCA and its associated methods, are certain to have increasing influence on how major sections of our diagnostic manuals are structured in the future.

Survival/Hazard Analysis

Whereas latent class analysis works well with an array of dichotomous indicator variables to describe samples in cross-section, other variables of interest to diagnostic research are neither dichotomous nor continuously scaled, but are rather *temporal* or time-dependent. That is, they have to do with the age of onset of an illness, the amount of time that elapses between the acquisition of the first symptom and the next, the length of latency to relapse after treatment, and so on. A family of techniques called *event-history analyses* or, more commonly, *survival/hazard analysis* (Cox & Oakes, 1984; Singer & Willett, 1991), can be applied to data such as these.

Basically, survival/hazard methods model the temporal pattern by which a group of respondents changes from one state (e.g., unaffected and symptom free) to another state (e.g., symptomatic or, more formally, diagnosable). These methods have the uncommon advantages of (1) being able to accommodate cases that have not yet experienced the outcome ("right-censored data"); (2) being able to show changes across time, rather than cross-sectionally; and (3) producing intuitive graphical plots of the data that may highlight obscure relationships (Langenbucher & Chung, 1995). Survival/hazard methods are therefore powerful tools for the nosologist, as well as for more applied clinical scientists, who may use survival/hazard methods to study relapse risk, to compare outcomes for different groups over time, and for many other purposes. In addition, advanced survival/hazard methods permit scientists to simultaneously input mediating or moderating variables that influence the survival function, just as covariates or moderating/mediating variables are used in more common multivariate routines.

Survival/hazard methods require the experimenter to define several parameters. First is the *anchor event,* the initial terminus of the *observation window* that "starts the clock," after which the *failure* or *index event* (such as the first occurrence of a symptom) can occur at any time for any particular respondent. That is, at least until the *censoring event* defines the final terminus of observation. The essence of the method is the analysis of the interval between the anchor event and the failure event or censoring event (whichever comes first), using dichotomous data (the failure has/has not occurred) for each discrete period of time during the observation window, until either the failure event occurs (rendering that case a failed or affected case) or the censoring event occurs (rendering that case "right-censored"). This ability of survival/hazard analysis to utilize data from right-censored cases makes it unique and powerful, because all other methods for describing temporality (e.g., mean time to relapse, comparing group means for time to relapse for two different treatment groups) use data only from fully expressed (uncensored) cases. This is because censored cases, those that have not yet experienced the index event, have an indeterminate value on the temporality variable, and so cannot be factored into the mean or average group value, though they may be

less severe, or different in some other way, from cases in which the index event unfolds quickly. Thus, if the observation window is short and a large proportion of subjects are censored, only event-history analysis can take their information into account.

The basic product of event-history analysis is the "survival curve," based on a mathematical function that estimates the proportion of cases at each point in time after the anchor event that will have survived the onset of the failure event. The logarithmic transformation of the survival function is the *hazard function,* which estimates the probability of succumbing to the index or failure event during each discrete interval in the observation window. The most common technique of estimating the survival/hazard function uses Kaplan-Meier statistics to develop negatively decelerating survival curves decreasing from 100 percent survival at the anchor event to ever lower percentages surviving as the observation window marches forward. Survival curves with steep slopes at the early time points describe subject groups for whom the index event tends to occur relatively early; more gradual slopes indicate subject groups that tend to survive the onset of the event for a longer period. The log rank test and Wilcoxon tests are commonly used statistics for testing for between-group differences in survival/hazard functions, during the early and late periods of the observation window, respectively.

In an early application, Burke, Burke, Rae, and Regier (1991) used survival/hazard analysis to test whether major depression has become a serious concern for individuals at younger and younger ages—a matter of clinical, policy, and even social import. Using data from the ECA, the authors examined age of onset for unipolar depression (as well as for bipolar disorder, three classes of anxiety disorders, and alcohol/drug abuse and dependence) in four birth cohorts—6,566 individuals born before 1917; 4,432 individuals born between 1917 and 1936; 4,981 individuals born between 1937 and 1952; and 4,766 individuals born from 1953 to 1966—using survival/hazard methods. The anchor event was the respondent's birth, the failure event was first diagnosis of depression (or the other comparison disorders) in the respondent's lifetime as assessed by the DIS (the measure used by ECA interviewers), and the censoring event was age 30, selected by the researchers because of their interest in the early onset of depression. Separate survival/hazard functions were run for each birth cohort, and for each of the disorders tested. The data showed that there has indeed been a gradual shift across birth cohort to younger ages of onset for major depression, in particular increases in hazard in the late teens and twenties. Similar shifts were observed in age of onset for alcohol and drug abuse/dependence in the most recent cohorts, particularly those that came of age in the 1960s and '70s. Interestingly, there were no systematic cohort effects observed in the other comparison categories—bipolar disorder and anxiety disorders. The authors used the complex set of temporal analyses to speculate on social conditions as well as psychological variables that in the latter half of the twentieth century induced earlier and earlier onset of major depression and substance use disorders in Americans. The study is widely cited, and served as one of the sparks for an increasing attention to early-onset, even childhood, depression, in U.S. psychiatry.

IRT Analysis. Classical test theory (CTT), which had a pronounced effect on nosological methods in the 1970s and '80s, is now gradually being supplanted by item response theory (IRT; Embretson & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1991) in the field of test development, though it has yet to see much employment in diagnostic research. That will likely change. Whereas IRT shares with CTT the assumption that a latent construct (such as *DSM-IV* alcohol dependence) is tapped by a set of items (such as *DSM-IV* criteria for alcohol dependence), it appears to have distinct advantages over CTT for understanding the behavior of diagnostic criteria. Because IRT models can provide a much more detailed picture of how well a criterion set functions than can methods based on CTT, their application to diagnostic measures should prove more and more valuable.

IRT assumes that a single latent construct is tapped by a set of items. This typically requires the preliminary step of confirmatory factor analysis to ". . . provide supporting 'evidence' that a data set is reasonably dominated by a single common factor" (Embretson & Reise, 2000). In diagnostic research, this requires administering a diagnostic interview like the SCID to a large sample of persons with varying degrees of illness or illness risk, noting for each subject which diagnostic criteria are met and which are not met, then exposing the group data to a confirmatory factor analysis routine like Mplus (Muthen & Muthen, 1998) to make sure that the diagnostic criteria tap only a single dimension of psychopathology (called a *latent trait* in the common IRT parlance). Binary IRT models can then be tested. The two-parameter model is most relevant to the situation in which psychiatric symptoms are queried dichotomously (symptom present/absent) by structured interview.

The two-parameter model run with IRT methods such as MULTILOG (Thissen, 1991) obtains two measures of the performance of each criterion. The first is the criterion's *threshold:* the point on the underlying dimension of psychopathology at which 50 percent of respondents endorse the item (i.e., report that they have the symptom). Threshold is therefore related to *item difficulty* in CTT. The second mea-

sure of a criterion's performance, *discrimination,* is an indicator of the degree of precision with which the criterion can distinguish between subjects with higher versus lower levels of psychopathology.

IRT methods offer several kinds of graphical output of interest to nosologists. An *item response curve* (IRC) is an S-shaped logistic function that shows the probability of endorsing the criterion at each level of the underlying psychopathology. Both threshold and discrimination are contained in the shape of the IRC, with low threshold criteria being located far to the left on the dimension of underlying psychopathology, high threshold criteria being located far to the right; and with poorly discriminating criteria showing a "lazy" or gradually ascending S-shaped curve, and highly discriminating criteria showing an abrupt, steep slope to the IRC. Each IRC can be transformed into an *item information curve* (IIC), in most cases a roughly bell-shaped curve that indicates the amount of psychometric information provided by the criterion at all points along the dimension of psychopathology. When criteria are combined into a common measure of the dimension of psychopathology, as they typically are in diagnostic practice and research, their IIC's are additive, and the functioning of the diagnostic criterion set as a whole can be indicated by its *test information curve* (TIC).

Once the information function of the criterion set is known, it is possible to examine how precisely the criterion set measures individuals at various ranges of the underlying dimension of psychopathology. If most or all criteria in a given set have the same or similar thresholds, the resulting TIC will be peaked, and the standard error plot will be U-shaped, indicating high measurement precision in a relatively narrow range of psychopathology (i.e., it is a *peaked test*). In comparison, if criterion thresholds are well distributed across the full range of psychopathology, the resulting test information curve will be relatively flat, indicating that measurement precision is more or less constant across the full range of the underlying dimension of psychopathology (i.e., it is an *equal-precise test,* and perhaps useful as a quasi-continuous measure of the latent trait).

In the first application of this type, Langenbucher et al. (2004) demonstrated the utility of IRT for evaluating the performance of criteria for diagnosing *DSM-IV* substance use disorders. Data on all 11 *DSM-IV* alcohol, cannabis, and cocaine symptoms were gathered from 372 adult addictions treatment subjects interviewed face-to-face with the CIDI-SAM. Mplus was used to test unidimensionality of the criteria, and MULTILOG was used to develop IRT parameters and graphical output. Two of the 11 criteria, "tolerance" and "legal problems," had very poor discrimination parameters (they failed to distinguish persons with many substance use symptoms from those with few symptoms) and fit poorly with a unidimensional factor model. Though they are included in the *DSM-IV* criteria for substance use disorders, they seem to measure something different from what the other criteria reflect, were dropped from subsequent analyses and should, the study argues, be dropped from any criterion list considered for *DSM-V,* an astonishing recommendation in the case of alcohol tolerance, given its central role in alcoholism assessment for nearly 60 years. The study showed that IRT can be used to identify diagnostic criteria with poor performance characteristics. Just as importantly, the study showed that IRT can be used to study the construct validity of *DSM-IV* diagnoses: The test information curves for the combined criteria showed only a single "peak" or area of precise information, suggesting that *DSM-IV* abuse and dependence criteria discriminate only "nondiagnosable" from "diagnosable" cases, with the best discrimination occurring at a moderate level of underlying psychopathology, equivalent to four criteria met (for alcohol), five met (for cannabis) or six met (for cocaine). This finding would appear to challenge the basic structure of this important set of *DSM-IV* categories, which purports to separate cases into undiagnosed, abuse, and dependence categories, with dependence commencing at a fairly mild level of involvement, any three or more of seven symptoms. IRT-based methods are obviously worthy of much more extensive application to symptom data, as the development process for *DSM-V* unfolds.

SUMMARY AND CONCLUSIONS

We have, in this chapter, explored the conceptual underpinnings of psychiatric diagnosis in the phenomena of *folk taxonomies* (Raven, Berlin, & Breedlove, 1971), *natural categories* (Rosch, 1973), and *prototype categorization* (Cantor et al., 1980). We developed the history of diagnosis from the earliest wonderings of Hippocrates, to Paracelsus's concept of *syndromal diagnosis,* to the contributions of Karl Kahlbaum and Emil Kraepelin, to the development of common but still naïve psychiatric nomenclatures in the years after World War II, all the way into a new empirical era with the arrival of the neo-Kraepelinian movement that delighted many while angering many of its critics (Rosenhan, 1973; Szasz, 1960). Along the way, we reviewed major findings on the reliability and validity of the diagnostic systems that have been constructed with such care and expense, on the development of new diagnostic interviews and how they have affected several branches of clinical science, particularly epidemiologic, treatment outcome, and descriptive psychopathology research. We introduced literature that is diverse, often conceptually

difficult, and too often impossible to be meaningfully accessed by nonspecialists. This is particularly true of the more advanced quantitative methods that are, even now, just beginning to make their impacts felt, methods like *survival analysis, ROC analysis, latent class analysis,* and methods based in Item Response Theory, but to which we wanted to introduce the reader.

Nothing in these pages should be clearer than this: Diagnosis and classification of psychiatric illness is an evolving craft. It is as old as is attention to mental illness itself—taking a prominent place in the writings of Hippocrates, Galen, Paracelsus, and many other ancient scholars and scientists—but it is at the same time, purely on the basis of its growth as an empirical science, a young science, a work in progress not much more than a quarter-century old, since the neo-Kraepelinians first put pen to paper to formulate new research methods (Feighner et al., 1972; Robins & Guze, 1970; Spitzer et al., 1975), to describe the results of their new interviews (Spitzer et al., 1967, 1970), to show the promise of their new diagnostic manuals (APA, 1980, 1987, 1994), and to cast new light on matters as various as the prevalence of post-traumatic stress disorder (Kessler et al., 1995), the age of onset of major depression (Burke et al., 1991), the most predictive symptoms of alcohol dependence (Langenbucher et al., 2004), and many other issues.

Many, many crucial questions still remain, particularly concerning the personality disorders, the permeability of boundaries between disorders of thought and disorders of affect regulation, the most parsimonious ways to *subtype* major classes of psychiatric illness, and many others. The research participants are there, the empirical methods for gathering data through structured interviews are there, as are exciting, powerful new methods to partition and explain the data. Even looking back over the road so far traveled, it is difficult to disagree with Spitzer and Williams (1987), that the development of psychiatric classification systems in the last half-century—in America, this has been the *DSM* tradition—grew into one of the most prestigious forces in the maturation of the mental health system generally. This is a trend we fully expect to continue and intensify as the next decades unfold.

REFERENCES

American Psychiatric Association. (1933). Notes and comments: Revised classified nomenclature of mental disorders. *American Journal of Psychiatry, 90,* 1369–1376.

American Psychiatric Association. (1952). *Diagnostic and statistical manual of mental disorders.* Washington, DC: Author.

American Psychiatric Association. (1968). *Diagnostic and statistical manual of mental disorders* (2nd ed.). Washington, DC: Author.

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: Author.

American Psychiatric Association. (1987). *Diagnostic and statistical manual of mental disorders* (3rd ed., rev.). Washington, DC: Author.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Andreasen, N. C. (1995). The validation of psychiatric diagnosis: New models and approaches. *American Journal of Psychiatry, 152,* 161–162.

Baldessarini, R., Finkelstein, S., & Arana, G. (1983). The predictive power of diagnostic tests and the effect of prevalence of illness. *Archives of General Psychiatry, 40,* 569–573.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4,* 561–571.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J. E., & Erbaugh, J. K. (1962). Reliability of psychiatric diagnoses: II. A study of consistency of clinical judgments and ratings. *American Journal of Psychiatry, 119,* 351–357.

Black, S. (1971). Labeling and psychiatry: A comment. *Social Science and Medicine, 5,* 391–392.

Blashfield, R. K. (1984). *The classification of psychopathology.* New York: Plenum Press.

Blashfield, R. K., & Draguns, J. G. (1976). Toward a taxonomy of psychopathology. *British Journal of Psychiatry, 42,* 574–583.

Brown, T. A., Di Nardo, P. A., Lehman, C. L., & Campbell, L. A. (2001). Reliability of *DSM-IV* anxiety and mood disorders: Implications for the classification of emotional disorders. *Journal of Abnormal Psychology, 110,* 49–58.

Burke, K. C., Burke, J. D., Rae, D. S., & Regier, D. A. (1991). Comparing age of onset of major depression and other psychiatric disorders by birth cohorts in five US community populations. *Archives of General Psychiatry, 48,* 789–795.

Canino, G., Shrout, P. E., Rubio-Stipec, M., Bird, H. R., Bravo, M., Ramirez, R., et al.(2004). The *DSM-IV* rates of child and adolescent disorders in Puerto Rico. *Archives of General Psychiatry, 61,* 85–93.

Cantor, N., Smith, E. E., French, R. deS., & Mezzich, J. (1980). Psychiatric diagnosis as prototype categorization. *Journal of Abnormal Psychology, 89,* 181–193.

Caplan, P. J. (1991). How do they decide who is normal? The bizarre, but true, tale of the *DSM* process. *Canadian Psychology, 32,* 162–70.

Carson, R. C. (1991). Dilemmas in the pathway of the *DSM-IV. Journal of Abnormal Psychology, 100,* 302–7.

Cassidy, F., & Carroll, B. J. (2001). The clinical epidemiology of pure and mixed manic episodes. *Bipolar Disorders, 3,* 35–40.

Cole, D. A. (1987). Utility of confirmatory factor analysis in test validation research. *Journal of Consulting and Clinical Psychology, 55,* 584–594.

Compton, W. M., & Guze, S. B. (1995). The neo-Kraepelinian revolution in psychiatric diagnosis. *European Archives of Psychiatry and Clinical Neuroscience, 245,* 196–201.

Corbitt, E. M., & Widiger, T. A. (1995). Sex differences among the personality disorders: An exploration of the data. *Clinical Psychology: Science and Practice, 2,* 225–238.

Cox, D. R., & Oakes, D. (1984). *Analysis of survival data.* London: Chapman and Hall.

Dilsaver, S. C., Chen, R., Shoaib, A. M., & Swann, A. C. (1997). Phenomenology of mania: Evidence for distinct depressed, dysphoric, and euphoric presentations. *American Journal of Psychiatry, 156,* 426–430.

Dollfus, S., Everitt, B., Ribeyre, J. M., Assouly-Besse, F., Sharp, C., & Petit, M. (1996). Identifying subtypes of schizophrenia by cluster analyses. *Schizophrenia Bulletin, 22,* 545–555.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Endicott, J., & Spitzer, R. L. (1978). A diagnostic interview: The Schedule for Affective Disorders and Schizophrenia. *Archives of General Psychiatry, 35,* 837–844.

Faraone, S. V., & Tsuang, M. T. (1994). Measuring diagnostic accuracy in the absence of a "gold standard." *American Journal of Psychiatry, 151,* 650–657.

Feighner, J. P., Robins, E., Guze, S. B., Woodruff, R. A., Winokur, G., & Munoz, R. (1972). Diagnostic criteria for use in psychiatric research. *Archives of General Psychiatry,* 26, 57–63.

Fennig, S., Kovasznay, B., Rich, C., Ram, R., Pato, C., Miller, A., et al. (1994). Six-month stability of psychiatric diagnoses in first-admission patients with psychosis. *American Journal of Psychiatry, 151,* 1200–1208.

Frances, A. J., Widiger, T. A., & Pincus, H. A. (1989). The development of *DSM-IV. Archives of General Psychiatry, 46,* 373–375.

Garmezy, N. (1978). Never mind the psychologists: Is it good for the children? *The Clinical Psychologist, 31,* 1–6.

Goldman, H. H., Rye, P., & Sirovatka, P. (1999). *Mental health: A report of the surgeon general.* (GPO No. 2004082500218). Washington, DC: U.S. Department of Health and Human Services.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Hartung, C. M., & Widiger, T. A. (1998). Gender differences in the diagnosis of mental disorders: Conclusions and controversies of *DSM-IV. Psychological Bulletin, 123,* 260–278.

Hasin, D. S., Trautman, K. D., Miele, G. M., Samet, S., Smith, M., & Endicott, J. (1996). *Psychiatric research interview for substance and mental disorders (PRISM).* New York: Research Assessment Associates.

Haslam, N., & Beck, A. T. (1994). Subtyping major depression: A taxometric analysis. *Journal of Abnormal Psychology, 103,* 686–692.

Helzer, J. E., Clayton, P. J., Pambakian, R., Reich, T., Woodruff, R. A., & Reveley, M. A. (1977). Reliability of psychiatric diagnosis: II. The test/retest reliability of diagnostic classification. *Archives of General Psychiatry, 34,* 136–141.

Helzer, J. E., Robins, L. N., Taibleson, M., Woodruff, R. A., Reich, T., & Wish, E. D. (1977). Reliability of psychiatric diagnosis: I. Methodological review. *Archives of General Psychiatry, 34* (Suppl. 328), 18–21.

Hollingshead, A. B., & Redlich, F. C. (1955). Social mobility and mental illness. *American Journal of Psychiatry, 112,* 179–185.

Houts, A. C. (2002). Discovery, invention, and expansion of the modern *Diagnostic and statistical manual of mental disorders.* In L. E. Beutler & M. L. Malik (Eds.), *Rethinking the DSM: A psychological perspective* (pp. 17–65). Washington, DC: American Psychological Association.

Johnson, J. G., Cohen, P., Kotler, L., Kasen, S., & Brook, J. S. (2002). Psychiatric disorders associated with risk for the development of eating disorders during adolescence and early adulthood. *Journal of Consulting and Clinical Psychology, 70,* 1119–1128.

Jöreskog, K., & Sörbom, D. (1989). *LISREL 7: Users reference guide.* Mooresville, IN: Scientific Software, Inc.

Kendell, R., & Jablensky, A. (2003). Distinguishing between the validity and utility of psychiatric diagnoses. *American Journal of Psychiatry, 160,* 4–12.

Kendler, K. S. (1990). Toward a scientific psychiatric nosology: Strengths and limitations. *Archives of General Psychiatry, 47,* 969–73.

Kendler, K. S., Eaves, L. J., Walters, E. E., Neale, M. C., Heath, A. C., & Kessler, R. C. (1996). The identification and validation of distinct depressive syndromes in a population-based sample of female twins. *Archives of General Psychiatry, 53,* 391–399.

Kendler, K. S., McGuire, M., Gruenberg, A. M., & Walsh, D. (1995). Examining the validity of *DSM-III-R* schizoaffective disorder and its putative subtypes in the Roscommon Family Study. *American Journal of Psychiatry, 152,* 755–764.

Kendler, K. S., Neale, M. C., & Walsh, D. (1995). Evaluating the spectrum concept of schizophrenia in the Roscommon Family Study. *American Journal of Psychiatry, 152,* 749–754.

Kendler, K. S., & Roy, M.-A. (1995). Validity of a diagnosis of lifetime major depression obtained by personal interview versus family history. *American Journal of Psychiatry, 152,* 1608–1614.

Kessler, R. C., Avenevoli, S., & Merikangas, K. R. (2001). Mood disorders in children and adolescents: An epidemiologic perspective. *Biological Psychiatry, 49,* 1002–1014.

Kessler, R. C., Sonnega, A., Bromet, E., Hughes, M., & Nelson, C. B. (1995). Posttraumatic stress disorder in the National Comorbidity Survey. *Archives of General Psychiatry, 52,* 1048–1060.

Kessler, R. C., Stein, M. B., & Berglund, P. (1998). Social phobia subtypes in the National Comorbidity Survey. *American Journal of Psychiatry, 155,* 613–619.

Kim-Cohen, J., Caspi, A., Moffitt, T. E., Harrington, H., Milne, B. J., & Poulton, R. (2003). Prior juvenile diagnoses in adults with mental disorder. *Archives of General Psychiatry, 60,* 709–717.

Kirk, S. A., & Kutchins, H. (1992). *The selling of DSM: The rhetoric of science in psychiatry.* Hawthorne, NY: Walter de Gruyter.

Klein, D. N., Ouimette, P. C., Kelly, H. S., Ferro, T., & Riso, L. P. (1994). Test-retest reliability of team consensus best-estimate diagnoses of Axis I and Axis II disorders in a family study. *American Journal of Psychiatry, 151,* 1043–1047.

Klerman, G. L., Vaillant, G. E., Spitzer, R. L., & Michels, R. (1984). A debate on *DSM-III. American Journal of Psychiatry, 141,* 539–553.

Koenigsberg, H. W., Harvey, P. D., Mitropoulou, V., Schmeidler, J., New, A. S., Goodman, M., et al. (2002). Characterizing affective instability in borderline personality disorder. *American Journal of Psychiatry, 159,* 784–788.

Langenbucher, J. W., & Chung, T. (1995). Onset and staging of *DSM-IV* alcohol dependence using mean age and survival/hazard methods. *Journal of Abnormal Psychology, 104,* 346–354.

Langenbucher, J., Labouvie, E., Sanjuan, P., Kirisci, L., Bavly, L., Martin, C., et al. (2004). Item Response Theory analysis of alcohol, cannabis and cocaine criteria in *DSM-IV. Journal of Abnormal Psychology, 113,* 72–80.

Lilienfeld, S. O., & Marino, L. (1995). Mental disorder as a Roschian concept: A critique of Wakefield's "harmful dysfunction" analysis. *Journal of Abnormal Psychology, 104,* 411–420.

Mataix-Cols, D., Rauch, S. L., Baer, L., Eisen, D. M., Goodman, W. K., Rasmussen, S. A., et al. (2002). Symptom stability in adult obsessive-compulsive disorder: Data from a naturalistic two-year follow-up study. *American Journal of Psychiatry, 159,* 263–268.

Mattanah, J. J. F., Becker, D. F., Levy, K. N., Edell, W. S., & McGlashan, T. H. (1995). Diagnostic stability in adolescents followed up 2 years after hospitalization. *American Journal of Psychiatry, 152,* 889–894.

McCutcheon, A. L. (1987). *Latent class analysis.* Thousand Oaks, CA: Sage.

Minsky, S., Vega, W., Miskimen, T., Cara, M., & Escobar, J. (2003). Diagnostic patterns in Latino, African American, and European American psychiatric patients. *Archives of General Psychiatry, 60,* 637–644.

Morey, L. C. (1988). The categorical representation of personality disorder: A cluster analysis of *DSM-III-R* personality features. *Journal of Abnormal Psychology, 97,* 314–321.

Mota, V. L., & Schachar, R. J. (2000). Reformulating attention-deficit/hyperactivity disorder according to signal detection theory. *Journal of the American Academy of Child & Adolescent Psychiatry, 39,* 1144–1151.

Murphy, J. M., Berwick, D. M., Weinstein, M. C., & Borus, J. F. (1987). Performance of screening and diagnostic tests: Application of receiver-operating characteristic analysis. *Archives of General Psychiatry, 44,* 550–555.

Muthen, B., & Muthen, L. (1998). *Mplus.* Los Angeles: Muthen.

Nathan, P. E. (1967). *Cues, decisions, and diagnoses.* New York: Academic Press.

Nathan, P. E. (1998). *DSM-IV* and its antecedents: Enhancing syndromal diagnosis. In J. W. Barron (Ed.), *Making diagnosis meaningful: New psychological perspectives* (pp. 3–27). Washington, DC: APA Books.

Nathan, P. E., Andberg, M. M., Behan, P. O., & Patch, V. D. (1969). Thirty-two observers and one patient: A study of diagnostic reliability. *Journal of Clinical Psychology, 25,* 9–15.

Nathan, P. E., & Harris, S. L. (1980). *Psychopathology and Society* (2nd ed.). New York: McGraw-Hill.

Nathan, P. E., & Langenbucher, J. W. (1999). Psychopathology: Description and classification. In J. T. Spence (Ed.), *Annual Review of Psychology* (Vol. 50, pp. 79–107). Palo Alto, CA: Annual Reviews.

Nelson, C. B., Little, R. J. A., Heath, A. C., & Kessler, R. C. (1996). Patterns of *DSM-III-R* alcohol dependence symptom progression in a general population survey. *Psychological Medicine, 26,* 449–460.

Nelson, E., & Rice, J. (1997). Stability of diagnosis of obsessive-compulsive disorder in the Epidemiologic Catchment Area study. *American Journal of Psychiatry, 154,* 826–831.

Peralta, V., & Cuesta, M. J. (2003). The nosology of psychotic disorders: A comparison among competing classification systems. *Schizophrenia Bulletin, 29,* 413–425.

Prigerson, H. G., Shear, M. K., Jacobs, S. C., Reynolds, C. F. III, Maciejewski, P. K., Davidson, J. R. T., et al. (1999). Consensus criteria for traumatic grief: A preliminary empirical test. *British Journal of Psychiatry, 174,* 67–73.

Raven, P. H., Berlin, B., & Breedlove, D. E. (1971). The origins of taxonomy. *Science, 174,* 1210–1213.

Regier, D. A., Myers, J. K., Kramer, M., Robins, L. N., Blazer, D. G., Hough, R. L., et al. (1984). The NIMH Epidemiologic Catchment Area program: Historical context, major objectives and study population characteristics. *Archives of General Psychiatry, 41,* 934–941.

Robbins, J. M., Kirmayer, L. J., & Hemami, S. (1997). Latent variable models of functional somatic distress. *Journal of Nervous & Mental Disease, 185,* 606–615.

Robins, E., & Guze, S. (1970). Establishment of diagnostic validity in psychiatric illnesses: Its application to schizophrenia. *American Journal of Psychiatry, 126,* 983–987.

Robins, L. N., Helzer, J. E., Croughan, J., & Ratcliff, K. S. (1981). National Institute of Mental Health Diagnostic Interview Schedule: Its history, characteristics, and validity. *Archives of General Psychiatry, 38,* 381–389.

Robins, L. N., Wing, J., Wittchen, H.-U., Helzer, J. E., Babor, T. F., Burke, J. D., et al. (1988). The Composite International Diagnostic Interview: An epidemiologic instrument suitable for use in conjunction with different diagnostic systems and in different cultures. *Archives of General Psychiatry, 45,* 1069–1077.

Rosch, E. R. (1973). Natural categories. *Cognitive Psychology, 4,* 328–350.

Rosenhan, D. L. (1973). On being sane in insane places. *Science, 179,* 250–258.

Sanislow, C. A., Grilo, C. M., Morey, L. C., Bender, D. S., Skodol, A. E., Gunderson, J. G., et al. (2002). Confirmatory factor analysis of *DSM-IV* criteria for borderline personality disorder: Findings from the Collaborative Longitudinal Personality Disorders Study. *American Journal of Psychiatry, 159,* 284–290.

Schacht, T. E. (1985). *DSM-III* and the politics of truth. *American Psychologist, 40,* 513–526.

Schacht, T. E., & Nathan, P. E. (1977). But is it good for the psychologists? Appraisal and status of *DSM-III. American Psychologist, 32,* 1017–1025.

Schneider, K. (1959). *Clinical Psychopathology.* New York: Grune & Stratton.

Schuckit, M. A., Smith, T. L., Danko, G. P., Bucholz, K. K., Reich, T., & Bierut, L. (2001). Five-year clinical course associated with *DSM-IV* alcohol abuse or dependence in a large group of men and women. *American Journal of Psychiatry, 158,* 1084–1090.

Shea, M. T., Stout, R., Gunderson, J., Morey, L. C., Grilo, C. M., McGlashan, T., et al. (2002). Short-term diagnostic stability of schizotypal, borderline, avoidant, and obsessive-compulsive personality disorders. *American Journal of Psychiatry, 159,* 2036–2041.

Simpson, S. G., McMahon, F. J., McInnis, M. G., MacKinnon, D. F., Edwin, D., Folstein, S. E., et al. (2002). Diagnostic reliability of Bipolar II disorder. *Archives of General Psychiatry, 59,* 736–740.

Singer, J., & Willett, J. (1991). Modeling the days of our lives: Using survival analysis when designing and analyzing longitudinal studies of duration and timing of events. *Psychological Bulletin, 110,* 268–290.

Speilberger, C. S., Gorsuch, R. L., & Lushene, R. E. (1970). *Manual for the State-Trait Anxiety Inventory.* Palo Alto, CA: Consulting Psychologists Press.

Spitzer, R. L. (1983). Psychiatric diagnosis: Are clinicians still necessary? *Comprehensive Psychiatry, 24,* 399–411.

Spitzer, R. L., & Endicott, J. (1968). DIAGNO: A computer program for psychiatric diagnosis utilizing the differential diagnostic procedure. *Archives of General Psychiatry, 18,* 746–756.

Spitzer, R. L., & Endicott, J. (1969). DIAGNO II: Further developments in a computer program for psychiatric diagnosis. *American Journal of Psychiatry, 125,* 12–21.

Spitzer, R. L., Endicott, J., Fleiss, J. L., & Cohen, J. (1970). The Psychiatric Status Schedule: A technique for evaluating psychopathology and impairment of role functioning. *Archives of General Psychiatry, 23,* 41–55.

Spitzer, R. L., Endicott, J., & Robins, E. (1975). *Research diagnostic criteria (RDC) for a selected group of functional disorders.* New York: New York State Psychiatric Institute.

Spitzer, R. L., Fleiss, J. L., Endicott, J., & Cohen, J. (1967). Mental Status Schedule: Properties of a factor-analytically derived scale. *Archives of General Psychiatry, 16,* 479–493.

Spitzer, R. L., & Williams, J. B. W. (1982). The definition and diagnosis of mental disorders. *Sage Annual Reviews of Studies in Deviance, 6,* 15–31.

Spitzer, R. L., & Williams, J. B. W. (1986). *Structured clinical interview for DSM-III.* New York: New York State Psychiatric Institute, Biometrics Research Department.

Spitzer, R. L., & Williams, J. B. W. (1987). Revising *DSM-III:* The process and major issues. In G. L. Tischler (Ed.), *Diagnosis and classification in psychiatry: A critical appraisal of DSM-III* (pp. 425–434). New York: Cambridge University Press.

Spitzer, R. L., Williams, J. B. W., Gibbon, M., & First, M. B. (1992). The structured clinical interview for *DSM-III-R* (SCID). I: History, rationale, and description. *Archives of General Psychiatry, 49,* 624–629.

Spitzer, R. L, Williams, J. B. W., & Skodol, A. E. (1980). *DSM-III:* The major achievements and an overview. *American Journal of Psychiatry, 137,* 151–164.

Srole, L., Langer, T. S., Michael, S. T., Opler, M. K., & Rennie, T. A. (1962). *Mental health in the metropolis: The Midtown Manhattan Study.* New York: McGraw-Hill.

Strauss, J. S. (1975). A comprehensive approach to psychiatric diagnosis. *American Journal of Psychiatry, 132,* 1193–1197.

Szasz, T. S. (1960). The myth of mental illness. *American Psychologist, 15,* 113–118.

Thissen, D. (1991). *MULTILOG user's guide—Version 6.* Chicago: Scientific Software, Inc.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58,* 236–244.

Widiger, T. A., Frances, A. J., Pincus, H. A., First, M. B., Ross, R., & Davis, W. (Eds.). (1994). *DSM-IV sourcebook* (Vol. 1). Washington, DC: American Psychiatric Association.

Widiger, T. A., Frances, A. J., Pincus, H. A., Ross, R., First, M. B., & Davis, W. (Eds.). (1996). *DSM-IV sourcebook* (Vol. 2). Washington, DC: American Psychiatric Association.

Widiger, T. A., Frances, A. J., Pincus, H. A., Ross, R., First, M. B., & Davis, W. (Eds.). (1997). *DSM-IV sourcebook* (Vol. 3). Washington, DC: American Psychiatric Association.

Widiger, T. A., Frances, A. J., Pincus, H. A., Ross, R., First, M. B., Davis, W., et al. (Eds.). (1998). *DSM-IV sourcebook* (Vol. 4). Washington, DC: American Psychiatric Association.

Widiger, T. A., & Trull, T. J. (1993). The scholarly development of *DSM-IV.* In J. A. C. e Silva & C. C. Nadelson (Eds.), *International review of psychiatry* (pp. 59–78). Washington, DC: American Psychiatric Press.

Wittchen, H.-U., Zhao, S., Kessler, R. C., & Eaton, W. W. (1994). *DSM-III-R* generalized anxiety disorder in the National Comorbidity Survey. *Archives of General Psychiatry, 51,* 355–364.

World Health Organization. (1967). *International classification of diseases* (8th ed.). Geneva, Switzerland: Author.

World Health Organization. (1992). *International classification of diseases and related health problems* (10th rev.). Geneva, Switzerland: Author.

Yen, S., Shea, T., Pagano, M., Sanislow, C. A., Crilo, C. M., McGlashan, T. H., et al. (2003). Axis I and Axis II disorders as predictors of prospective suicide attempts: Finding from the Collaborative Longitudinal Personality Disorders study. *Journal of Abnormal Psychology, 112,* 375–381.

Zilboorg, G. (1941). *A history of medical psychology.* New York: Norton.

Zubin, J. (1967). Classification of behavior disorders. *Annual Review of Psychology, 18,* 373–406.

CHAPTER 2

Research Considerations: Latent Variable Approaches to Studying the Classification and Psychopathology of Mental Disorders

LAURA CAMPBELL-SILLS AND TIMOTHY A. BROWN

INTRODUCTION

Research in psychopathology covers a wide and varied range of content and methods. Perusal of any journal devoted to the study of psychopathology demonstrates this diversity: readers encounter articles on epidemiology, genetics, diagnosis and classification, psychophysiology, neuroimaging, development of assessment tools, experimental investigations of behavior, and many other topics. The research methods common to these content areas diverge substantially, and each subspecialty faces its own set of challenges in developing rigorous research design and data analytic approaches.

We have chosen to devote this chapter on research considerations to examining a methodological approach that has broad applicability to psychopathology research: structural equation modeling (SEM). Although SEM is not relevant to all domains of psychopathology research (e.g., it cannot be applied in small sample investigations), it has proven useful for evaluating wide-ranging hypotheses about the classification, course, and etiology of psychological disorders. Accurate classification, description, and causal models provide a foundation for any science, yet many questions remain regarding how to best classify and explain psychopathological phenomena. In this chapter, we present SEM as a "state-of-the-art" method for seeking answers to these crucial questions.

In the sections that follow, we briefly describe the procedures and advantages of SEM before moving to an overview of guidelines for maximizing some of this methodology's most important features (e.g., evaluation of model fit, modeling measurement error). We also explicate applications of SEM that have particular relevance for psychopathology but tend to be underutilized in applied research (e.g., construct validation techniques, measurement invariance testing). Finally, we identify common problems encountered with different SEM applications and recommend strategies for minimizing their impact. Readers are encouraged to refer to resources cited throughout the chapter for further examples and for explanations of technical aspects of SEM applications.

OVERVIEW OF STRUCTURAL EQUATION MODELING

Structural equation modeling (SEM) is increasingly used to address key questions in psychopathology research. In a recent introduction to a special section on SEM in the *Journal of Abnormal Psychology,* Tomarken and Baker (2003) reviewed the diverse areas that have been covered in this premiere psychopathology journal using SEM. Topics included (but were not limited to) assessment of construct validity, tests of etiological models, evaluation of genetic and environmental contributions to disorders, and modeling the course of symptoms over time. It is difficult to conceive of any other methodology currently in use that covers such vast conceptual territory.

SEM allows researchers to define latent constructs of interest (e.g., depression, neuroticism, impulsivity) by multiple, observable indicators (e.g., questionnaire scores, interviewer ratings, physiological responses). The direct relationships specified between observed measures (indicators) and their corresponding latent variables constitute the *measurement* component of an SEM model (along with the estimates and relationships among the indicator error variances; i.e., "error theory"). SEM also can be utilized to test causal models of the relationships among latent variables. The hypothesized relations among latent variables represent the *structural* component of an SEM model. Psychopathology research focused on construct or questionnaire validation might focus exclusively on evaluation of measurement models; whereas studies

of etiological models of psychological disorders would also incorporate tests of structural models.

SEM can accommodate considerably more complex models of psychopathology than alternative data analytic methods (e.g., analysis of variance [ANOVA], multiple regression). For example, a single SEM analysis can incorporate multiple outcome variables and different types of nontautological relationships (e.g., indirect effects). Most constructs of interest in psychopathology research (e.g., psychological disorders) are conjectured to have multiple determinants (e.g., genetics, temperament, environmental factors) and to impact a variety of outcomes (e.g., work functioning, relationships). Therefore, SEM's capacity to simultaneously evaluate a complex array of relationships among multiple predictor and outcome variables allows researchers to reproduce the relationships among psychopathological constructs with greater verisimilitude.

SEM offers numerous other advantages over more traditional statistical approaches. For example, SEM helps psychopathology researchers surmount one of the major limitations of most investigations: imperfect measurement. Virtually no constructs studied in the field of psychopathology are free from measurement error. Nevertheless, traditional data analytic methods assume perfect reliability of all measures (i.e., it is assumed that the observed score is equal to the "true" score). In contrast, SEM allows researchers to examine the relationships among constructs of interest while adjusting for the effects of measurement error. The most common SEM approach to addressing measurement error is the use of multiple indicators to define each latent variable. However, even variables measured by single indicators can be adjusted for measurement error in an SEM model (cf. Bollen, 1989). Moreover, researchers may specify an error theory (e.g., modeling correlated error) that allows their models to approximate true relationships among variables even more closely.

Another benefit of SEM is the availability of statistics that assess the "goodness of fit" of the hypothesized model. In instances where the model is "overidentified" (i.e., the number of freely estimated parameters in the model is less than the number of elements—e.g., variances, covariances—of the input matrix), goodness-of-fit indexes provide a statistical evaluation of how well the hypothesized model accounts for the observed relationships in the data. The most common form of SEM analysis generates a predicted covariance matrix based on the resulting parameter estimates of the specified model (e.g., estimates from a path model A → B → C, would estimate the relationship between A and C as the product of the A → B and B → C paths). The predicted covariance matrix is compared to the sample (observed) covariance matrix. The residual covariance matrix reflects the discrepancies between the predicted and observed relationships (e.g., the observed relationship between A and C versus the relationship between A and C that is predicted by the A → B → C path model). In various manners, goodness-of-fit statistics summarize the degree of discrepancy between the observed and predicted matrices.

Another advantage of SEM is the ability to statistically compare hypothesized and competing models. Researchers are able to specify a model based on prior empirical evidence and theory, evaluate how well this hypothesized model fits the data, and compare its fit to that of plausible alternative models. This process improves upon the more common practice in psychopathological research in which models are assumed to be valid if they achieve statistical significance. With judicious use of SEM, a researcher can reasonably assert that his or her hypothesis produced a good-fitting model in which statistically significant relationships were observed, and that this model explained the data better than potentially viable alternative models.

As with any statistical technique, the utility of SEM depends on its proper application by psychopathology researchers. For example, not all models are testable in SEM. In order to be mathematically "identified," the hypothesized model must contain the same or a fewer number of freely estimated parameters (e.g., factor loadings, regressive paths) than the number of elements in the input matrices (e.g., indicator variances, covariances). If this number is equal, the model is "just-identified" (has zero *df,* where *df* = number of elements of input matrix minus the number of freely estimated parameters), and goodness-of-fit evaluation does not apply (i.e., the model can be tested, but goodness of fit is ensured to be perfect).[1] As noted previously, goodness of fit evaluation applies to overidentified solutions (i.e., models associated with *df*s ≥ 1). "Underidentified" models cannot be tested in SEM. A solution may be underidentified mathematically (i.e., the number of estimated parameters exceeds the size of the input matrix, $df < 0$) or empirically (i.e., the model has $df \geq 1$, but aspects of the specification and/or input matrix preclude the solution from obtaining a unique or valid set of parameter estimates). Complex model specifications (e.g., models with a large number of correlated errors, double-loading indicators, or bidirectional relationships) may often result in statistical or empirical underidentification.

Moreover, the sample data must satisfy the mathematical assumptions of the fitting function selected for the SEM analysis. The fitting function is the statistical algorithm used to estimate the model's parameter estimates (and standard errors) to minimize the difference between the observed and predicted matrices. The vast majority of SEM analyses are conducted with maximum likelihood (ML), the default esti-